16.30 – Proyecto Final de Carrera

PROYECTO FINAL

Desarrollo de herramientas informáticas para la construcción de Gold Standards de señales fisiológicas

Estudiante:

- *Cordara, Camila – Legajo: 55327*
-

Tutor:

- *Prof. Dr. Ing. Madorno, Matias*

MBMED
CREATING MEDICAL SOLUTIONS

ITBA

Agradecimientos

Al Instituto Tecnológico de Buenos Aires, por brindarme un entorno para desarrollarme personal y académicamente.

A MBMED, por facilitar los recursos y herramientas para la realización de este Proyecto Final de Carrera.

A el Prof. Dr. Ing. Matias Madorno, por su ayuda y buena predisposición en rol de tutor.

Al Dr. Pablo Fierens, por su colaboración en la búsqueda bibliográfica y sus sugerencias.

A todos los profesionales de la salud que aportaron su tiempo y conocimientos para la realización de este Proyecto Final de Carrera.

Por último, a mis amigos y familia, que me acompañaron y apoyaron durante el transcurso de la carrera.

Índice

Glosario

ACV: ventilación controlada

EAdi: actividad eléctrica del diafragma

EPOC: enfermedad pulmonar obstructiva crónica

EPOC: enfermedad pulmonar obstructiva crónica

GAS: Google Apps Script

I:E: relación inspiración/espiración

IA: índice de asincronías

ICC: correlaciones intraclase

ISMD : dispositivo de medición mal estructurado

NIV: ventilación no invasiva

Paw: presión de la vía aérea

PCV/AV: ventilación asistida-controlada por presión

PCV: ventilación controlada por presión

Pe: proporción esperada de acuerdo

PEEP: presión espiratoria final positiva

Peso: presión esofágica

Po: proporción observada de acuerdo

PSV: ventilación con presión de soporte

SDRA: síndrome de distrés respiratorio agudo

SIMV: ventilación intermitente mandatoria sincronizada

Te: tiempo espiratorio

Ti: tiempo inspiratorio

Ti-M: tiempo de insuflación de la máquina

Ti-N: tiempo de inspiración neuronal

UCI: unidad de cuidados intensivos

VCV/AC: ventilación asistida-controlada por volumen

VCV: ventilación controlada por volumen

VM: ventilación mecánica

VMI: ventilación mecánica intermitente

Vt: volumen corriente

Resumen

En la actualidad hay un creciente interés en el desarrollo de algoritmos inteligentes de apoyo a la toma de decisiones médicas debido a que permiten la reducción de los errores médicos, hacen los tratamientos más costo-efectivos y aumentan la seguridad del paciente. Estos sistemas de apoyo a la toma de decisiones se basan en distintos softwares que realizan clasificación automática de señales fisiológicas en categorías de interés y la detección automática de patologías o signos relevantes en esas señales. Sin embargo, para poder validar el funcionamiento de estos últimos softwares es necesario comparar sus resultados con un Gold Standard y en diversas áreas de la medicina, aún no hay disponibilidad de estos Gold Standards.

Es de utilidad desarrollar una gran cantidad de Gold Standards: para diferentes señales fisiológicas y, a su vez, para cada tipo de señal de acuerdo con diferentes categorías, patologías y grupos poblacionales. Por eso, la herramienta desarrollada en este Trabajo Final tiene como objetivo facilitar la elaboración de Gold Standards, mediante la automatización de varias etapas del proceso para generar bases de señales nomencladas.

Debido a la amplitud de uso de la herramienta, se decidió mostrar su funcionamiento y utilidad mediante un ejemplo de uso. Se utilizó la herramienta para generar una de base señales respiratorias nomencladas de asincronías paciente-ventilador para pacientes de UCI con VM y SDRA ya que es un área en la cual sería relevante contar con un Gold Standard para mejorar los resultados clínicos.

La plataforma desarrollada, en base a Google Apps Script, fue útil para generar una encuesta totalmente cruzada con 100 respiraciones a clasificar. Los únicos datos de entrada que debió proveer el usuario son la base de señales fisiológicas y la clasificación. Se programó la modificación de un software para generar imágenes a partir de señales respiratorias, el cual puede ser utilizado por un usuario sin conocimientos de informática y podría ser aplicable en cualquier otro proceso de generación de Gold Standard para señales respiratorias.

En el ejemplo de uso, los resultados obtenidos a partir de la información recolectada (es decir, las clasificaciones de los evaluadores) fueron muy alentadores ya que se mostró un elevado nivel de

acuerdo entre los expertos. En base a un análisis descriptivo, existió un muy alto nivel de consenso en 89% de las respiraciones analizadas. Considerando la totalidad de respiraciones (100), el acuerdo porcentual fue de 81,9% y tanto el kappa de Fleiss como el kappa de Light fueron de 0,70. Si se consideran únicamente las 89 respiraciones en las que hubo un acuerdo superior al 71,4% de los evaluadores, el acuerdo porcentual fue de 88,0% y el kappa de Davies y Fleiss y el kappa de Light, de 0,79 y 0,78 respectivamente.

Todos los valores obtenidos indican un acuerdo sustancial y para el cual se pueden establecer conclusiones tentativamente. Como era de esperar, si se consideran únicamente las 89 respiraciones con alto consenso, los valores del acuerdo porcentual y de ambas estadísticas kappa se acercan a ser considerados acuerdo perfecto y para los cuales se pueden establecer conclusiones definitivamente. Estos resultados muestran que diferentes estadísticas dan resultados coherentes que indican que el acuerdo es elevado en todos los casos, ya sea considerando las 100 respiraciones totales o las 89 en las que hay alto consenso. Por lo tanto, esto muestra la validez de los datos recolectados en la encuesta a los expertos para ser utilizados en la construcción de un Gold Standard, lo cual evidencia la utilidad de la herramienta desarrollada para el caso de asincronías.

La experiencia de los usuarios de la herramienta de generación de encuestas fue buena dado que fue encontrada como sencilla y rápida de usar. En conjunto con los buenos resultados de la encuesta de experiencia de usuario a los expertos evaluadores, el análisis de experiencia de usuario muestra el potencial de adopción de la herramienta en grupos de profesionales de la salud.

Los buenos resultados de consenso y la satisfactoria experiencia de usuario muestran la utilidad y el potencial del sistema desarrollado, que puede aplicarse no sólo para el área respiratoria sino para cualquier señal fisiológica. Estos Gold Standards pueden contribuir a optimizar el funcionamiento de los sistemas de soporte de decisiones y del avance de la medicina digital. Por su parte, la herramienta desarrollada puede facilitarles el camino a los investigadores en estas nuevas tecnologías.

Introducción a la problemática

En la actualidad las innovaciones tecnológicas están siendo disruptivas en todas las áreas y la industria de la salud no es una excepción. Cada vez existen más desarrollos centrados en algoritmos inteligentes que puedan brindar apoyo a la toma de decisiones médicas y es creciente el número de organizaciones de atención médica que los utilizan o se interesan en incorporarlos. Estos algoritmos están destinados a ayudar a los médicos en sus procedimientos de diagnóstico para tomar decisiones más precisas y efectivas, minimizando los errores médicos, mejorando la seguridad del paciente y haciendo los tratamientos más costo-efectivos.

La relevancia de este tipo de software de soporte a la decisión clínica ha llegado al punto que a fines de 2017, la Administración de Drogas y Alimentos (FDA), organismo regulatorio de Estados Unidos, publicó una guía en relación al camino regulatorio para las aplicaciones de software de salud digital [1].

El funcionamiento de estos sistemas de soporte de decisiones clínicas se basa en guías clínicas, sistemas de razonamiento basados en reglas y en casos, lógica difusa, redes neuronales, entre otras alternativas. Cualquiera de estos métodos requiere en primer lugar la disponibilidad de sistemas que permitan la clasificación automática de señales fisiológicas en categorías de interés y la detección automática de patologías o signos relevantes en esas señales. Dichos sistemas generan los datos de entrada en base a los cuales funcionan los sistemas de soporte para la toma de decisiones.

Existe un gran auge en el desarrollo de estos sistemas para clasificación y detección automática para señales fisiológicas en numerosas áreas de la medicina. Sin embargo, no siempre es posible evaluar su funcionamiento, es decir, existen diversos softwares de clasificación y detección automática que no han podido ser validados. Esto se debe a que para evaluar un sistema de este tipo se deben poder comparar sus resultados con un *benchmark*, es decir, con una base de señales nomencladas contra la cual se pueda realizar una validación. Dicho *benchmark* o Gold Standard no se encuentra disponible en varias áreas de la medicina, lo cual representa un gran obstáculo para la validación y mejoramiento de clasificación de señales fisiológicas y, en última instancia, de los sistemas de soporte a la toma de decisiones.

Existen bases de señales rotuladas para evaluar los algoritmos de clasificación y validar su funcionamiento. Una de estas es la de cardiología: hay una variedad de bases de datos de señales cardíacas en physionet.com, una iniciativa subvencionada por el National Institute of Health (NIH) de Estados Unidos y, en menor medida, por otras empresas como Medtronic. En PhysioNet las señales se encuentran nomencladas según diferentes tópicos de interés (por ejemplo: arritmia fetal, taquiarritmia ventricular espontánea). Sin embargo, como fue mencionado, en otras áreas de la medicina, aún no se cuenta con este tipo de bases de señales nomencladas y es necesario generarlas.

Siendo que hay una gran variedad de señales que es de interés clasificar para generar eventuales Gold Standards es de utilidad desarrollar de una herramienta que automatice el proceso de rotulado y cuantificación de datos para generarlos, la cual pueda adaptarse con facilidad a diferentes tipos de señales, categorías y distintos grupos poblacionales, sin tener que elaborar una herramienta específica para cada estudio particular.

En este Trabajo Final el objetivo es desarrollar una herramienta que permite simplificar el proceso de generación de Gold Standards, es decir, bases de señales nomencladas. La misma puede ser utilizada para obtener Gold Standard de cualquier señal fisiológica.

La ventilación mecánica es el procedimiento de soporte de vida más habitual en la Unidad de Cuidados Intensivos (UCI), por lo que la clasificación de señales respiratorias podría hacer una gran contribución al desarrollo de sistemas de soporte a la toma de decisiones clínicas más efectivos y es una de las áreas en las que no existen bases de datos de señales nomencladas aún. Para señales respiratorias se necesita contar con Gold Standards de acuerdo con varias clasificaciones, tales como los modos ventilatorios, el tipo de respiración (asistida, espontánea o mandatoria) para señales de respiración asistida, la cantidad de ruido cardíaco en las señales de presión esofágica, la existencia o no de asincronías, entre otros. A su vez, también es relevante hacer Gold Standards para distintos tipos de poblaciones o subgrupos (pacientes de UCI, pacientes con anestesia general, pacientes con ventilación invasiva, pacientes con SDRA, entre otros).

Las asincronías paciente-respirador son un tema particularmente relevante en medicina respiratoria dado que están asociadas a una mayor mortalidad [2] y a resultados deficientes de la VM [3],[4],[5].

Su identificación automática sería de gran utilidad para mejorar el tratamiento de los pacientes y equipamiento de la UCI.

En este informe, se muestra el funcionamiento y utilidad de la herramienta mediante el desarrollo de un Gold Standard de clasificación de señales respiratorias y, en particular, de asincronías paciente-respirador de pacientes con SDRA en las primeras 72 horas de ventilación mecánica. Se selecciona este subgrupo de pacientes por su alta tasa mortalidad, donde una pequeña mejora implica vidas que se salvan [5],[2].

Componentes del ciclo de una respiración

La respiración es el proceso fisiológico que permite que ocurra el intercambio gaseoso a nivel alveolar. Se considera que la respiración se divide en dos fases llamadas fase inspiratoria o inspiración y fase espiratoria o espiración. Durante la inspiración, el flujo de aire es hacia el interior del cuerpo. La inspiración en el caso de VM puede ser disparada por el respirador o por el paciente y el gas es entregado por el respirador (es un proceso activo). Al final de la inspiración, el flujo de gas cesa y la respiración se transforma en espiración. Durante esta fase, el sistema respiratorio elimina el dióxido de carbono y otros gases.

En VM, el punto de transición de la fase inspiratoria a la fase espiratoria se denomina habitualmente "ciclado". La fase inspiratoria pasa a la espiración cuando el flujo de gas cesa del ventilador mecánico y comienza el flujo espiratorio. Muchos ajustes en el ventilador mecánico inducen el ciclo, como el volumen, el tiempo y el flujo preestablecidos. Los recientes avances en el diseño de ventiladores mecánicos permiten a los médicos desempeñar un papel más importante en la evaluación y manipulación de los ciclos [6].

Se pueden ver las formas de las señales de volumen, flujo, presión de la vía aérea y presión esofágica, en la Figura 1, durante la inspiración (fondo color gris claro) y la espiración (fondo color gris oscuro).

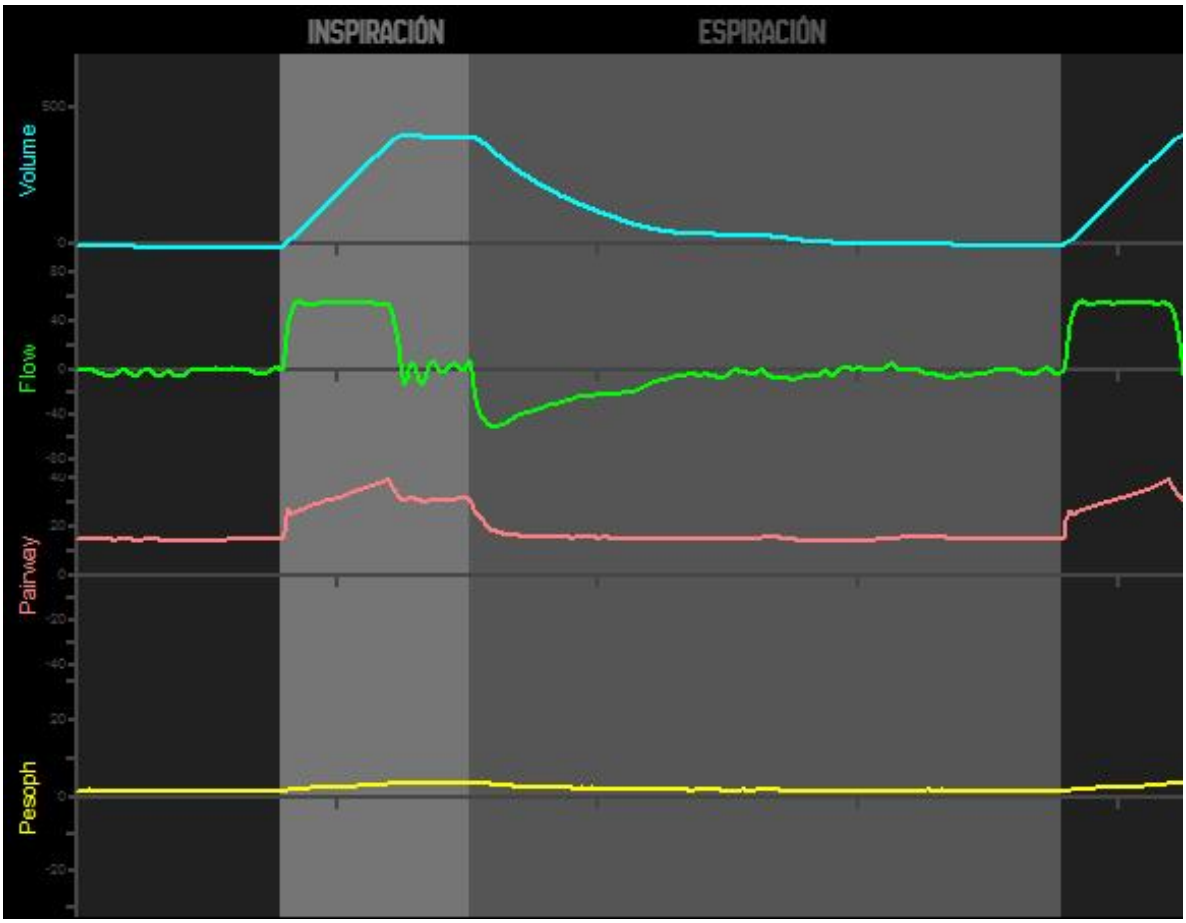


Figura 1. Se pueden ver las formas de las señales de volumen (en celeste), flujo (en verde), presión de la vía aérea (en rojo) y presión esofágica (en amarillo), durante una ventilación. Se distinguen la inspiración por fondo color gris claro y la espiración por el fondo color gris oscuro). La imagen fue tomada del archivo Centro médico 3, y es la respiración número 163.

Ventilación mecánica y asincronías paciente-respirador

La ventilación mecánica es el tratamiento de soporte vital más común en la UCI. Entre sus finalidades, se encuentra reducir o reemplazar el trabajo de respiración del paciente y mejorar el intercambio de gases [2],[5],[7].

Las variables que definen la asistencia que brinda el ventilador son el disparo, el ciclado y la variable objetivo del sistema de control. El disparo se refiere al momento en el que el ventilador comienza a asistir al paciente entregándole aire, es decir, marca el inicio de la respiración. Por su lado, ciclado

refiere al fin de la asistencia del ventilador y, por lo tanto, de la inspiración, a partir de lo cual el paciente espirará en forma pasiva.

Existen diversos modos ventilatorios, de acuerdo con el patrón de flujo, volumen y presión en la vía aérea del aire entregado al paciente, y actualmente no hay un consenso para definir el modo ventilatorio utilizado para tratar a los pacientes. La elección del modo ventilatorio se toma teniendo en cuenta la patología respiratoria o condición que llevó al paciente a requerir asistencia respiratoria mecánica, el carácter agudo o crónico, el patrón ventilatorio, el estado hemodinámico del paciente, entre otros factores. De hecho, en el caso de un mismo paciente cuya situación clínica y fisiopatológica evoluciona durante su estadía hospitalaria, es necesario adaptar la configuración del ventilador regularmente.

Una clasificación habitual de los modos ventilatorios los divide en dos grandes grupos, de acuerdo con la variable objetivo del mecanismo del control: volumen o presión. En los modos volumétricos, un volumen específico de gas es entregado (el Vt configurado). Por ende, Vt es constante, se debe seleccionar la forma de onda del flujo y el tiempo inspiratorio. La variable resultante es la presión, que será la necesaria para generar el flujo. La presión puede variar de respiración a respiración, y depende para cada paciente de la *compliance* y resistencia de la vía aérea. Por su parte, en los modos que controlan en base a la presión, se configura una presión inspiratoria de la vía aérea. Por lo tanto, la variable controlada por el ventilador es la presión de la vía aérea y alveolar, mientras que Vt puede variar de respiración a respiración siendo que la entrega de flujo de aire es variable según las características mecánicas del paciente.

Los modos utilizados en la ventilación mecánica han ido evolucionando a lo largo del siglo XX, a medida que progresó la medicina. Entre los hitos claves para comprender el avance en el conocimiento de ventilación mecánica, se encuentra la introducción de la ventilación intermitente mandatoria (VMI) en 1973, realizado por Downs et al, como un método para facilitar el destete en adultos [8]. Esta nueva concepción de compartir el trabajo respiratorio entre el paciente y el ventilador, denominada usualmente soporte ventilatorio parcial, fue adoptada rápidamente como método para mecánica ventilatoria debido a sus múltiples beneficios. Sin embargo, conlleva también la aparición de la asincronía paciente-ventilador [9].

En la medicina respiratoria actual, de acuerdo con la situación clínica de los pacientes, se puede optar por una ventilación mecánica que suprima por completo el trabajo respiratorio de los pacientes o bien una ventilación empleada como soporte parcial. Los modos donde el esfuerzo del paciente no interviene se utilizan cuando este esfuerzo puede ser injuriante o en anestesia. Por su lado, la ventilación como soporte parcial es el modo más habitual debido a que ofrece numerosas ventajas, como una menor necesidad de sedación, mejor oxigenación y menor riesgo de deterioro hemodinámico y de atrofia muscular respiratoria [10],[5]. Además, el paciente controlado va a pasar a un modo asistido si no fallece antes. No obstante, las asincronías pueden socavar estos beneficios y, por lo tanto, constituyen un tópico de gran interés y relevancia clínica en el ámbito de medicina crítica.

La interacción entre el paciente y ventilador mecánico involucra múltiples factores. Interviene la interacción entre características del paciente (la hiperinflación y la fuerza muscular), en conjunto con la configuración del ventilador (por ejemplo, el criterio de disparo y de ciclado), es decir, la asistencia que provee el mismo.

Asumiendo que el sistema respiratorio se comporta como un modelo de dos elementos, uno con un comportamiento resistivo y otro con comportamiento elástico, a los cuales se le suma la presión muscular se puede deducir la ecuación de movimiento.

La ecuación de movimiento se define como:

$$Pmúsculo + Pventilador = flujo\ x\ resistencia + volumen\ x\ elastancia + PEEP$$

No se conoce en detalle el funcionamiento del drive respiratorio que genera la presión muscular, pero sí se sabe que está modulado por varios factores que funcionan como mecanismos fisiológicos de control. Entre ellos, se destacan el mecánico (efecto del volumen en la presión de los músculos), el químico (quimiorreceptores como gases sanguíneos y pH), el de reflejos (Hering-Breuer) y el de comportamiento (régimen de sedación). A su vez, la conjunción de estos mecanismos define la respuesta de los músculos respiratorios (la presión de los músculos) y el tiempo neural de inspiración y espiración [5].

Un *drive* respiratorio alto puede deberse a un aumento de las demandas metabólicas o un intercambio de gases alterado y/o estímulos mecánicos intensos a través de receptores de pulmón.

Por otro lado, un *drive* respiratorio bajo se puede atribuir a un sistema nervioso central deprimido, por sedación (excesiva) y/o un soporte ventilatorio excesivo.

Una administración apropiada de ventilación mecánica como soporte ventilatorio parcial reduciría la insuficiencia respiratoria del paciente al disminuir el trabajo respiratorio excesivo, mientras se mantiene un nivel adecuado de esfuerzo espontáneo. Esto puede ser alcanzado mediante una interacción armoniosa paciente-ventilador. Sin embargo, es usual que este equilibrio no pueda lograrse con éxito completo y exista una discordancia entre las necesidades del paciente y la asistencia otorgada por el respirador. Las asincronías paciente-ventilador ocurren tanto en el contexto de un *drive* respiratorio alto (por lo general asociado a una asistencia insuficiente) o un *drive* respiratorio bajo (generalmente relacionado con la sobre resistencia).

Se suele definir como asincronía paciente-respirador como el evento de una falta de coincidencia entre los tiempos inspiratorios y espiratorios propios del paciente y la entrega por parte del respirador mecánico. Las asincronías son un fenómeno complejo de caracterizar, identificar y tratar dado que involucran la interacción del respirador con varios órganos: los pulmones, músculos respiratorios, incluido el diafragma, y el sistema nervioso, incluidos los centros respiratorios [5].

En las últimas dos décadas, la tecnología y la investigación clínica han permitido un análisis más detallado de la interacción paciente respirador. El interés en las asincronías paciente-ventilador puede ser demostrada por el aumento notable en la cantidad de publicaciones realizadas en los últimos años [11], como ilustra la Figura 2.

Es relevante detectar las asincronías para mejorar los resultados del tratamiento de los pacientes y su confort durante la ventilación mecánica, dado que el impacto negativo de las asincronías paciente-respirador es de público conocimiento desde hace más de dos décadas. Una de las primeras observaciones relativa a los efectos perjudiciales de las asincronías paciente-respirador fue realizada por Chao et al en 1997 [12] refiriéndose al extendido tiempo de destete de aquellos pacientes con gran incidencia de asincronías (en particular, su estudio se centró en un tipo de asincronías denominadas esfuerzos inefectivos o *ineffective efforts*). En la actualidad, se ha demostrado la existencia de una clara asociación entre las asincronías y resultados deficientes de VM como: disfunción diafragmática o de músculos respiratorios inducida por el ventilador, mayor

tasa de traqueotomía, mayor duración de la VM [13] y una mayor mortalidad [12], [2], [14]. Aún no se ha demostrado científicamente si la relación entre la alta incidencia de asincronía y el empeoramiento de los resultados es causal o simplemente asociativa [11], pero aún en el caso que sea asociativa, poder obtener más información sobre las asincronías para buscar modos de reducirlas es relevante.

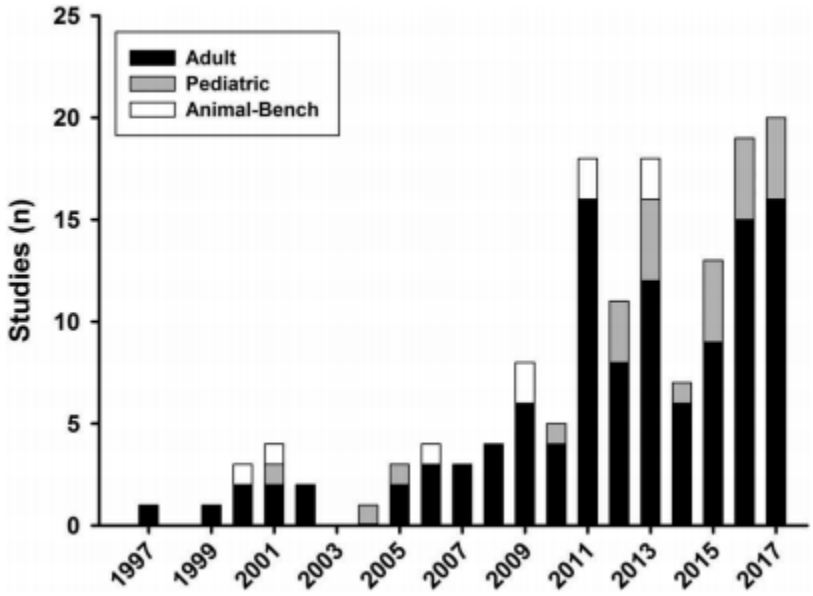


Figura 2. Estudios publicados sobre asincronía paciente-ventilador desde 1997 hasta 2017. Los estudios se muestran en conjunto y se dividen de acuerdo con los sujetos evaluados, es decir, adulto (negro), pediátrico (gris) y animal-banco (blanco). La cantidad de estudios publicados en pacientes adultos y pediátricos ha aumentado notablemente en la última década. Esta Figura fue tomada del artículo "Recognizing, quantifying and managing patient-ventilator asynchrony in invasiva and noninvasive ventilation" de Garófalo et al. [11].

Si bien existen actualmente diversos estudios para obtener datos epidemiológicos sobre las asincronías dado que aún no se conoce la medida exacta en que la asincronía paciente-ventilador afecta la duración de la ventilación mecánica, la disfunción muscular y la mortalidad, se trata de un evento frecuente en pacientes ventilados, como se puede ver en la Tabla 1 [15], [4], [9], [14], [16]. En la investigación realizada por Thille et al [9], los autores informan que durante la ventilación asistida invasiva se produce una alta incidencia de asincronías paciente-ventilador en el 25% de los

pacientes, que se caracterizan por peores resultados. En este mismo sentido, de acuerdo con estimaciones realizadas por Telias et al [5] al menos un tercio de los pacientes presenta una asincronía frecuente durante la ventilación mecánica. De acuerdo a otro estudio realizado por Garófalo et al [11] las asincronías clínicamente relevantes están presentes en hasta el 25% de los pacientes sometidos a ventilación mecánica invasiva y en hasta el 60% de los que reciben ventilación no invasiva.

En la Tabla 1, se muestra una comparación de una serie de estudios destacados que denotan la prevalencia de las asincronías durante ventilación mecánica. Se utiliza un índice de asincronías global (IA) como método de cuantificar las asincronías. El IA se define como el número de eventos asincrónicos dividido la frecuencia respiratoria total (esta última incluye tanto ciclos ventilatorios como esfuerzos inefectivos), expresado en porcentaje. Se considera que un índice de asincronías grave es mayor que 10% [17] y se encuentra asociado a peores resultados [11]. En un estudio realizado por Blanch et al [2], se ha demostrado que comparando pacientes con IA mayor 10% en comparación con aquellos que tuvieron IA menor a 10%, a pesar de haber tasas similares de reintubación y traqueostomía, hubo mayor mortalidad hospitalaria y una tendencia hacia una mayor duración de la VM. Se debe destacar que el método de detección puede influir significativamente en el IA resultante. En aquellos casos en los que hay carencia señales adicionales que evalúen directamente el esfuerzo respiratorio de un paciente, como presión esofágica (Peso) o la actividad eléctrica del diafragma (EAdi), la tasa real de asincronía podría estar potencialmente subestimada [18].

La reducción de la asincronía mediante la adaptación de las condiciones ventilatorias y de sedación es factible. Por lo tanto, aunque la asociación no prueba la causalidad, es de interés aumentar el conocimiento de las asincronías para buscar formas de mejorar los resultados clínicos al reducir o eliminar las asincronías paciente-ventilador. Por lo tanto, la detección de asincronías es crucial para reducir su incidencia, mediante ajustes de la configuración del ventilador, el régimen de sedación y otros factores técnicos y clínicos [11].

Citación	Número de participantes	Condición de los pacientes	Modo ventilatorio	Señales medidas	Porcentaje de pacientes con índice de asincronías > 10
Chao at al [15] 1997	174	VM prolongada	VCV/AC	-Flujo -Presión de la vía aérea -Presión esofágica (7 pacientes consintieron inserción de catéter esofágico)	10,9% (AI = 45% +/- 13,8%)
Thille et al [9] 2006	62	Ventilación invasiva en UCI superior a 24 horas	VCV/AC, PSV	-Flujo -Presión de la vía aérea	24%
Wit M [14] 2009	60	Ventilación mecánica en UCI durante las primeras 24 horas	VCV/AC o PCV/AC, SIMV+PSV, PSV	-Flujo -Presión de la vía aérea	27%
Vignaux [4] 2009	60	Falla respiratoria aguda (falla respiratoria	Modo NIV - Ventilador UCI	-Flujo -Presión de la vía aérea	43%

		crónica aguda, neumonía, post extubación, edema pulmonar, post operatorio, trauma torácico)		-Electromiografía diafragmática superficial	
Vignaux [16] 2010	65	Falla respiratoria aguda (falla respiratoria crónica aguda, neumonía, post extubación, edema pulmonar, post operatorio, trauma torácico)	Modo NIV - Ventilador UCI	-Flujo -Presión de la vía aérea -Electromiografía diafragmática superficial	38%
Robinson et al [17] 2013	35	Trauma	SIMV, PSV	-Flujo -Presión de la vía aérea -Volumen	25,7% (AI = 10,86%)

Tabla 1. Comparación de una serie de estudios que denotan la prevalencia de las asincronías durante ventilación mecánica. Se tienen en cuenta modos ventilatorios invasivos y no invasivos para demostrar la prevalencia de las asincronías (en particular, este proyecto se limita a pacientes con ventilación invasiva). Los distintos modos ventilatorios se explican en la sección Ventilación Mecánica.

La detección de las asincronías en la actualidad requiere como mínimo un análisis detallado de las formas de onda de flujo y presión de la vía aérea (Paw) que están disponibles en la pantalla del ventilador, junto con la evaluación clínica del patrón de respiración del paciente en particular. En otras palabras, actualmente la detección depende en primer lugar que el médico clínico a cargo esté

disponible para observar la pantalla atentamente y de forma continua. Además, incluso en este caso, la experiencia del especialista en el área de detección de asincronías es imprescindible y la carencia de la misma reduce significativamente la sensibilidad [18]. En particular, se ha demostrado que el entrenamiento de los profesionales de la salud que tienen capacitación específica en ventilación mecánica aumentan su capacidad para identificar la asincronía mediante el análisis de formas de onda [19], [18].

Se han realizado diversos estudios con el fin de evaluar sistemáticamente el valor de la inspección visual de las formas de onda del ventilador en la detección de asincronías paciente-ventilador. Para pacientes con ventilación invasiva, se establece que la observación de la forma de onda del ventilador como único método no permite una evaluación adecuada de las asincronías [18]. Esto se debe a que, de acuerdo con el estudio realizado, la capacidad de los médicos de la unidad de cuidados intensivos para reconocer las asincronías fue en general bastante baja y disminuyó a mayor prevalencia. En cuanto a los pacientes sometidos a ventilación no invasiva, en un reciente estudio observacional [20], realizado en múltiples centros médicos, se muestra que la detección de asincronías por la inspección visual de la forma de onda del ventilador durante NIV es problemática y requiere una atención constante del médico a cargo.

En este contexto, es clara la importancia en la práctica clínica que tendría una herramienta automática de detección de asincronías. En los últimos años, emergió el desarrollo de softwares que graban automática y continuamente las formas de onda del ventilador [21]. Es un tema de interés en el cual también hay empresas multinacionales de la industria de la ventilación mecánica trabajando en soluciones de software para detección de asincronías, como General Electric, Medtronic, Getinge y Drager, y hay grupos de investigación enfocados en el tema, por ejemplo Pleural Pressure Working Group (PLUG) dirigido por Brochard, que fue creado como una división de la Sociedad Europea de Terapia Intensiva (ESCIM). En este contexto, la empresa nacional MBMed está desarrollando desde 2014 una solución de software dedicada para la detección automática de asincronías. Sin embargo, dicho software aún no ha podido ser validado debido a la ausencia de una base de datos de respiraciones clasificadas que sirva como Gold Standard.

Es por esto que de entre todas las bases de señales nomencladas que se podrían generar con la plataforma desarrollada, se decide utilizarla para generar un Gold Standard de clasificación de

asincronías. Este permitirá validar softwares de detección automática de asincronías, ya sea aquel desarrollado por MBMed u otro. Dichos softwares podrían contribuir a la realización de estudios clínicos para la recopilación de más información sobre los efectos de las asincronías y, en una segunda instancia, podrían permitir el desarrollo de sistemas de apoyo a la toma de decisiones clínicas que permitan minimizar la incidencia de las asincronías en los pacientes y mejorar los resultados de la ventilación mecánica.

Clasificación de asincronías

Las asincronías pueden clasificarse en diversas categorías o tipos. No existe un consenso absoluto entre los expertos en medicina respiratoria, pero se mencionan a continuación las categorías consideradas en el Gold Standard del presente trabajo, las cuales son las de uso más generalizado.

Ineffective Triggering

El *ineffective triggering* es un tipo de asincronía que también se conoce como *ineffective effort* o *wasted effort*. Se caracteriza por un esfuerzo inspiratorio del paciente que no logra disparar una respiración del ventilador, es decir, no es asistido por el ventilador.

Para identificarlo, se busca evidencia clínica de esfuerzo inspiratorio del paciente, sin una posterior insuflación mecánica. En las curvas, se puede localizar mediante una deflexión positiva de la curva de flujo (un aumento en el flujo durante la espiración), una deflexión negativa de Paw, y una deflexión negativa en Peso, no seguido de una respiración [5].

Puede ocurrir durante el ciclo mecánico inspiratorio o espiratorio y depender de diversos mecanismos, como un *drive* y/o esfuerzo respiratorio débil, una presión espiratoria positiva intrínseca alta (PEEPi) y un disparo del ventilador excesivamente bajo en sensibilidad [11].

Auto-triggering

El *auto-triggering* tiene lugar cuando el ventilador entrega asistencia no relacionada con el esfuerzo espontáneo del paciente. Ocurre cuando los cambios en la presión de la vía aérea y/o el flujo secundario a las oscilaciones cardíacas o fugas de aire se detectan erróneamente como esfuerzos de activación [9]. Por lo tanto, su aparición depende principalmente del tipo de disparo y la sensibilidad.

Short (or premature) cycling

Se produce *short cycling* o *premature cycling* cuando el esfuerzo inspiratorio del paciente continúa durante la espiración mecánica. El tiempo de inspiración neuronal del paciente (Ti-N) es mayor que la insuflación real administrada por el ventilador (Ti-M).

La activación de los músculos inspiratorios durante la deflación mecánica da como resultado una contracción excéntrica del diafragma, que es potencialmente perjudicial para los músculos respiratorios [5].

El *premature cycling* es más frecuente en pacientes con baja *compliance*, como los pacientes con SRDA [22], [23].

Cuando el *short cycling* es lo suficientemente fuerte puede desencadenar una segunda ventilación mecánica (*double triggering*) antes de la exhalación completa de la primera, lo que resulta en un aumento del volumen corriente total. Este último fenómeno también recibe el nombre de *breath stacking* debido a la aparición de una nueva respiración sobre la anterior.

Double triggering

El *double triggering* se puede caracterizar en sentido amplio por dos ciclos mecánicos activados por el paciente, separados por un muy corto tiempo de expiración (inferior al 30% del tiempo inspiratorio medio) [9]. Se produce porque la respiración mecánica termina antes de que se complete el esfuerzo del paciente que, después de una breve fase de exhalación, desencadena una segunda respiración mecánica.

El *double triggering* suele producirse cuando los pacientes con insuficiencia respiratoria reciben ventilación de soporte de presión (PSV). Un alto drive respiratorio contribuye al desarrollo de esta forma de asincronía [11].

Long/Delayed Cycling

El *long cycling* o *delayed cycling* es una condición en la que la asistencia mecánica excede la inspiración del paciente y se extiende hasta su propia espiración (neural). La insuflación mecánica continúa después de que la inspiración neuronal ha cesado, incluso durante la espiración activa.

Se puede detectar comparando la insuflación mecánica con la duración del esfuerzo inspiratorio usando Peso o EAdi [5].

Por ejemplo, durante PSV, el ventilador cicla cuando el flujo disminuye a un porcentaje establecido del flujo inspiratorio máximo. La insuflación tiende a ser más larga con niveles más altos de soporte de presión y con una mayor resistencia al flujo de aire. Además, los niveles más altos de soporte de presión resultan en un flujo máximo más alto que puede acortar el tiempo inspiratorio neural, lo que contribuye a una falta de coincidencia entre una insuflación mecánica prolongada y un corto tiempo de inspiración neural [5].

En general, los pacientes con enfermedad pulmonar obstructiva crónica (EPOC) y asma, que se caracterizan por una alta resistencia y una *compliance* pulmonar elevada, tienen esta asincronía con mayor frecuencia. El tiempo espiratorio más corto contribuye a empeorar la hiperinflación en estos pacientes [7].

Reverse triggering

Reverse triggering es un tipo de asincronía descripta recientemente y, si bien aún es de las menos estudiadas, ilustran una nueva forma de acoplamiento neuromecánico con consecuencias clínicas potencialmente importantes. Se trata de una condición en que la respiración es activada en sentido inverso, caracterizada por un esfuerzo muscular activado por el ventilador. Ocurre durante la ventilación mecánica controlada, con mayor incidencia en casos de sedación profunda y disminución del drive respiratorio. Una respiración es activada por el ventilador y es seguida por la actividad inspiratoria de los músculos respiratorios [5].

Este fenómeno se ha considerado como un *entrainment respiratorio* (o bloqueo de la fase respiratoria), en referencia a una relación temporal repetitiva fija establecida entre los ciclos respiratorios neurales y mecánicos

Entre los elementos que permiten identificar *reverse triggering* se debe mencionar una disminución en la Peso, un aumento en la EAdi después del comienzo de la insuflación mecánica, un aumento en el flujo espiratorio y una disminución de la Paw más adelante en el ciclo respiratorio.

Además, si el esfuerzo inspiratorio es lo suficientemente fuerte, el ventilador puede administrar una segunda respiración, lo que resulta en apnea (como el caso en drive respiratorio alto). La diferencia con el *double triggering* inducido por un drive alto (un *short cycling*) es que la primera respiración en el caso de la *reverse triggering* es una respiración mandatoria (no desencadenada por los pacientes).

La principal consecuencia del *reverse triggering* es que aumenta el Vt en pacientes. Clínicamente, es indeseable tener un Vt elevado porque existe una asociación entre altos Vt (12 mL/kg en comparación con 6 mL/kg) y una mayor mortalidad [2]. En este sentido, es importante destacar que no todas las asincronías son iguales: mientras algunas como *ineffective effort* se suelen considerar como de poca gravedad, el *reverse triggering* es de las más graves.

Plataforma de encuesta

El presente trabajo consistió en el desarrollo de un sistema que permite generar una encuesta o formulario a partir de archivos de señales fisiológicas y una clasificación en la que se desea rotular dichas señales. El objetivo final es automatizar el proceso de generación de Gold Standard, cuyas etapas se ilustran en la Figura 3. La herramienta generada automatiza la etapa de armado de encuesta de clasificación a partir de las imágenes generadas para cualquier tipo de señal fisiológica y de la definición de categorías por parte del usuario, mediante un *script*.

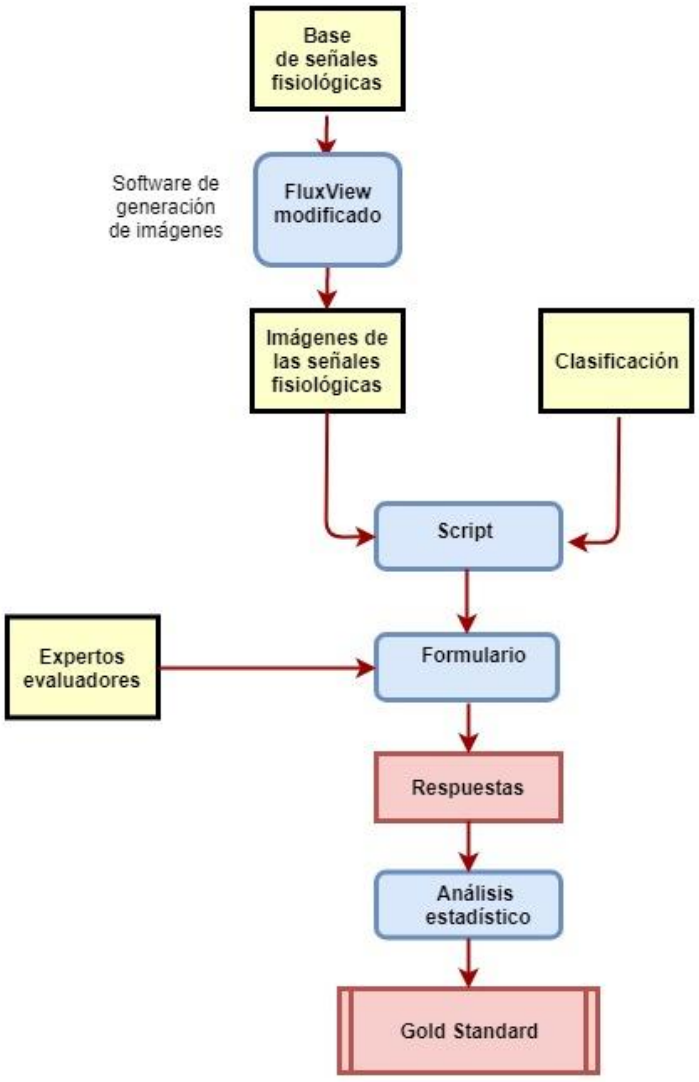


Figura 3. Diagrama de flujo mostrando los pasos esenciales necesarios para obtener un Gold Standard.

Siendo el sistema tan amplio, se consideró que la mejor forma de mostrar su potencial es a través de un ejemplo de uso. En las siguientes secciones se muestra la utilización de la herramienta que facilita el desarrollo de bases de señales rotuladas en la generación de un Gold Standards de señales respiratorias. Se seleccionó como criterio de clasificación de las respiraciones las asincronías paciente-ventilador y como población, a pacientes con SDRA durante las primeras 72 horas de ventilación mecánica.

Para este caso de ejemplo, se desarrolló un software que permite generar imágenes de señales respiratorias, que podría ser utilizado para cualquier otro caso de base de señales nomencladas respiratorias. Esto se encuentra representado con la herramienta "FluxView modificado" de la Figura 3.

En el proceso de desarrollo de la base nomencladas, se realizó primero una prueba preliminar con dos expertos y luego la clasificación definitiva con siete expertos. Esta prueba preliminar sirvió para involucrar a los expertos en la definición de los criterios de diseño para generar las imágenes y para la encuesta, que fueron considerados en el desarrollo del software de generación de imágenes ("FluxView modificado" en la Figura 3) y de la herramienta de generación de encuestas ("Script" en la Figura 3)

A lo largo del informe, se utiliza el término usuario para referirse al usuario de la herramienta para generar bases rotuladas, mientras que para hacer referencia a los usuarios finales que responden la encuesta se utilizan intercambiablemente los términos evaluadores, observadores, codificadores y expertos.

¿Por qué una plataforma de encuestas?

Actualmente no existe una base de respiraciones rotulada con asincronías paciente-respirador que sirvan como Gold Standard para validar algoritmos de identificación y detección automática. Únicamente se dispone de opiniones de profesionales de la salud expertos en medicina respiratoria, cuya precisión depende de su experiencia. Por lo tanto, para obtener una base de respiraciones nomencladas que funcionen como Gold Standard se debe consultar a diversos profesionales de la salud expertos en medicina respiratoria y en asincronías, para que clasifiquen un conjunto de

respiraciones y así poder analizar el consenso entre ellos. Para esto, se considera pertinente utilizar como herramienta una encuesta.

Siendo que no es posible utilizar una encuesta única para generar Gold Standard diferentes (para distintas señales fisiológicas, patologías y grupos poblacionales), el objetivo de este Trabajo es generar una herramienta que automatice el proceso de armado de encuestas de clasificación y el procesamiento de su información para simplificar la generación de bases de señales rotuladas.

Se busca desarrollar una herramienta digital de armado de encuestas, que pueda adaptarse fácilmente a cualquier tipo de señales fisiológicas y categorías. Además, una herramienta digital cuenta con el beneficio adicional de que la información que se obtenga para cada caso puede ser usada de forma directa para el cálculo de distintas variables de interés y estadísticas que permiten cuantificar el acuerdo de los expertos.

Criterios de diseño

Tipo de plataforma

Respecto a la plataforma para diseñar encuestas se evaluaron diferentes opciones, entre las cuales se seleccionó Google Forms, dado que dichos formularios pueden ser programados manualmente mediante Google Apps Script (basado en el lenguaje javascript). Esto permite tener una mayor flexibilidad y posibilidad de automatizar no solo la generación de encuesta, sino también el análisis de sus resultados, a diferencia de otros de los medios evaluados. También, Google Apps Script permite la interoperabilidad entre las distintas herramientas de Google. En este caso, carpetas de Google Drive (y su contenido), Google Spreadsheets, Google Forms y Gmail.

El Google Forms como modo de encuesta tiene numerosas ventajas. En primer lugar, al utilizar un entorno web, el formulario generado es apto para ser utilizado en una gran variedad de plataformas, desde computadoras de escritorio hasta celulares independientemente de su sistema operativo, lo cual fue considerado en función de mejorar la experiencia de usuario del experto que responde la encuesta. Además, el Google Forms tiene un *Spreadsheet* asociado donde se almacenan las respuestas de modo automático. Dicho *Spreadsheet* es útil para simplificar el posterior análisis

estadístico ya que permite que puedan exportarse los datos con facilidad e incluso procesarlos, dentro del mismo *Spreadsheet*, según las necesidades de cada usuario para cada Gold Standard.

Mediante Google Apps Script, se programó un *script* que permite crear un Google Forms, que es la encuesta que es enviada a los expertos para que ellos clasifiquen señales de interés. El *script* utiliza la información de un *Spreadsheet* como datos de entrada, el cual debe ser modificado por el usuario según los archivos de señales que desee nomenclar. El usuario puede elegir los elementos de interés de las señales fisiológicas a ser clasificados en dicho *Spreadsheet*, siendo la cantidad total variable dependiendo las necesidades de la encuesta que desee armar. Luego, puede ordenarlas aleatoriamente con el objetivo de que sean mostradas al evaluador experto en un orden al azar y minimizar un posible sesgo. Una vez que el usuario termina de modificar el *Spreadsheet* en base a la encuesta y Gold Standard que desea generar, debe definir las categorías de interés. Una vez que se ejecuta el *script*, se crea el Google Forms correspondiente (el cual será enviado a los especialistas) y un *Spreadsheet*, en el cual se almacenarán las respuestas y datos para su posterior procesamiento.

En el caso ejemplificador de clasificación de asincronías, las respiraciones fueron presentadas con imágenes que muestran varias señales respiratorias, las cuales se almacenan en una carpeta en Google Drive. Dichas imágenes fueron generadas mediante una modificación del código fuente del programa FluxView, que permite visualizar archivos de señales respiratorias y generar imágenes a partir de los mismos, ajustando las escalas y seleccionando respiraciones individuales. Como usuario, se hizo una elección de las respiraciones de interés en el *Spreadsheet* (en este caso, se trató de 100 respiraciones en total) y luego se las ordenó aleatoriamente. Por último, se definieron las categorías de asincronías a utilizar. Se puede elegir una cantidad variable de categorías de clasificación y en este caso se decidió utilizar diez categorías que representen las distintas asincronías.

Estructura de encuesta

En el caso del Gold Standard de asincronías paciente-respirador se utiliza la herramienta desarrollada con un diseño de encuesta totalmente cruzado (*fully-crossed*). Dicho tipo de diseño implica que todos los encuestados (es decir, todos los evaluadores expertos) deben clasificar todos los elementos seleccionados por el usuario como objetivo de clasificación. De esta forma, cada

elemento será clasificado por igual número de evaluadores y no se trata de una selección al azar entre un grupo más grande de evaluadores, sino que se trata siempre de los mismos (es decir, de todos los evaluadores involucrados).

Si bien este tipo de diseño tiene la desventaja principal de requerir un mayor número total de calificaciones, tiene numerosos beneficios tal como la posibilidad de evaluar y controlar el sesgo sistemático entre los codificadores en una estimación de IRR y así mejorar las estimaciones generales [24]. Se puede ver en la Tabla 2 una comparación entre un estudio totalmente cruzado y uno que no lo es.

En estudios centrados en el armado diseño metodológico de plataformas de clasificación [25], se distinguen entre estudios totalmente cruzados (*fully-crossed*), anidados (*nested*) y diseños de mediciones erróneamente estructurados (*ill-structured measurement designs*). En el caso de estudios anidados, hay dos alternativas de diseños. Si cada calificador califica un conjunto de objetos (respiraciones en este caso) único y no superpuesto, los objetivos se anidan dentro de los evaluadores. Alternativamente, si cada objetivo se califica mediante un conjunto de evaluadores único y no superpuesto, los evaluadores se anidan dentro de los objetivos. Por otro lado, para los diseños de mediciones erróneamente estructurados, los evaluadores y los sujetos no se cruzan por completo, ni se encuentran anidados. Estos estudios determinan que las medidas de concordancia (tal como las correlaciones intraclase o kappa) pueden subestimar la verdadera fiabilidad, a pesar de que abunda este tipo de diseño de encuestas en la investigación y la práctica que involucra la realización de calificaciones abundan.

Encuesta totalmente cruzada	**Evaluador 1**	**Evaluador 2**	**Evaluador 3**
Objeto 1	X	X	X
Objeto 2	X	X	X
Objeto 3	X	X	X

Encuesta no totalmente cruzada	Evaluador 1	Evaluador 2	Evaluador 3
Objeto 1	X	X	
Objeto 2		X	X
Objeto 3	X		X

Tabla 2. Tabla comparativa entre un diseño fully-crossed y uno que no lo es.

En particular, para el caso de encuesta de clasificación de asincronías, el diseño de encuesta totalmente cruzado (*fully-crossed*) implicó que todos los evaluadores (es decir, todos los expertos en medicina respiratoria) debieron clasificar la totalidad de las respiraciones analizadas. De esta forma, cada respiración fue clasificada por un número constante de evaluadores que fueron, de hecho, los mismos para todas las respiraciones.

Selección de señales

En primer lugar, antes de generar una encuesta, el usuario debe definir el Gold Standard que busca obtener, es decir, debe elegir qué señal fisiológica quiere clasificar y cuáles patologías desea utilizar como categorías. En base a dicha información, el usuario debe tomar una decisión respecto a cuáles señales necesita mostrar para realizar la clasificación y buscar una base de dichas señales que deberá ser utilizada como dato de entrada por la plataforma para armar la encuesta.

En el caso particular de la encuesta de asincronías respiratorias, dentro de todas las señales respiratorias posibles, aquellas imprescindibles para la detección de asincronías paciente-ventilador son la forma de onda de flujo, volumen y presión de la vía aérea, como ya se ha establecido en la literatura [3], [5], [13], [2]. Asimismo, se puede observar en la Tabla 1 que todos los estudios destacados incluyen ambas señales, aunque en algunos casos se incluyen también otras adicionales.

Hasta el 2010, se consideraba que no era excluyente disponer de la señal de presión esofágica para detección de asincronías, sino que era posible hacerlo en base al flujo y la señal de presión de la vía aérea. De acuerdo a Thille et al [3] en su estudio publicado en 2006, el número de asincronías

detectadas utilizando presión esofágica se encontraba correlacionado cercanamente con el número detectado usando señales de flujo y presión de la vía aérea.

No obstante, trabajos más recientes establecen que un monitoreo más avanzado usando la manometría esofágica o la actividad eléctrica del diafragma puede ayudar a los médicos a detectar asincronías, a pesar de que dicho monitoreo no se utiliza de forma rutinaria en la práctica clínica diaria debido a que es invasivo [5]. En particular, la señal de presión esofágica permite garantizar una buena estimación del esfuerzo de los músculos respiratorios del paciente [11] (es decir, el esfuerzo inspiratorio del paciente) y su adquisición es más factible en la práctica, en comparación con la actividad eléctrica del diafragma. Esto se debe a que en manometría esofágica se debe ingresar un balón de manometría hasta el esófago desde la nariz, mientras que para medir la actividad eléctrica del diafragma se deben utilizar electrodos de aguja (lo cual implica altos riesgos para la salud del paciente, como neumotórax) o bien una sonda de NAVA (Neurally Adjusted Ventilatory Assist) hasta el diafragma. La opinión de expertos en el área de mecánica ventilatoria [21] coincide en que las señales provistas por catéteres esofágicos (presión o electromiograma diafragmático) son los instrumentos más precisos y confiables para detectar asincronías y aumentan en gran medida la posibilidad de monitorear las asincronías [11]. Las señales esofágicas (ya sea que provean la actividad eléctrica del diafragma o presión esofágica) han permitido la descripción de la asincronía denominada *reverse triggering*.

En el caso de ejemplo de aplicación de la herramienta para desarrollar el Gold Standard de asincronías paciente-ventilador, se decidió utilizar tanto las señales de flujo y presión de la vía aérea, como la señal de presión esofágica. Se tuvo en consideración que los autores establecen que la presión esofágica tiene una gran contribución a la sensibilidad y especificidad en la detección de asincronías [5], [11], [18], [21]. Se debe destacar que, si el objetivo del Gold Standard fuera diferente, incluso en el área de señales respiratorias las señales elegidas podrían haber sido distintas (por ejemplo, flujo y presión de la vía aérea para clasificar si la respiración es mandatoria o no). El criterio de elección de las señales dependerá del objetivo del usuario que desea armar el Gold Standard.

Escala de tiempo para las imágenes

Usualmente, las señales fisiológicas se muestran en forma de curvas para poder ser clasificadas. Para cada Gold Standard, se deberá contar con un sistema que genere imágenes a partir de las señales fisiológicas en cuestión. Este software debe tener en cuenta la unidad y escala requerida para que en dicha imagen se pueda visualizar con claridad la señal fisiológica que se trate.

Para el caso ejemplificador de Gold Standard de clasificación de asincronías, se realizó una modificación del código fuente del programa FluxView para que genere imágenes de la representación de señales respiratorias en forma de curvas. Para definir los criterios de diseño, se consideraron las opiniones de los expertos consultados en una prueba preliminar.

En el proceso de modificación de dicho software, se debió definir qué amplitud de escala de tiempo (representada en el eje horizontal) utilizar para generar las imágenes. De acuerdo con los referentes consultados en la prueba preliminar es subóptimo tener una imagen de una respiración aislada para determinar si en la misma hay una asincronía y, en caso de que la hubiese, para realizar una clasificación correcta de la misma. En otras palabras, es de interés poder observar tanto la respiración en particular a ser clasificada como el patrón respiratorio del paciente (es decir, la visualización de las respiraciones anteriores y posteriores a la respiración en cuestión) para comparar cada respiración con sus adyacentes y, también, con el patrón respiratorio del paciente. En este mismo sentido, hay artefactos, como el movimiento peristáltico del esófago, que se observan mejor en la visualización de varias respiraciones sucesivas que en una sola respiración aislada. Además, existen tipos de asincronías relacionados entre sí que solo pueden apreciarse si se ven las respiraciones anteriores y posteriores a la que se debe clasificar. Por ejemplo, los *ineffective efforts* suelen encontrarse en clusters [26] o un *premature cycling* puede estar asociado a un *double triggering* posterior [5].

Finalmente, para este caso de ejemplo se decidió incluir para cada respiración dos imágenes: una imagen con zoom en el eje del tiempo de la respiración que se desea clasificar (Figura 4) y otra imagen de la tendencia del patrón respiratorio del paciente (Figura 5), es decir, que muestra en pantalla las respiraciones anteriores y posteriores. En ambas imágenes se seleccionó la respiración

que se desea clasificar de forma que sea clara la correlación entre las señales medidas para la misma respiración de manera vertical y la correspondencia entre ambas imágenes.

Selección del elemento a clasificar

Las señales fisiológicas se muestran en forma de curvas para poder ser clasificadas, como fue mencionado. Sin embargo, es necesario destacar en cada imagen el elemento a clasificar respecto de sus elementos adyacentes. Cada usuario que desee elaborar un Gold Standard debe elegir, cuando arme las imágenes correspondientes a sus señales fisiológicas, la manera de destacar el elemento a clasificar que le parezca más conveniente. Una de las técnicas más utilizadas es cambiar el color del fondo del elemento que se busca nomenclar para diferenciarlo del resto de la señal. Otra técnica es cambiar el color de las líneas de las curvas para el elemento en cuestión.

Es recomendable que para cada Gold Standard el usuario realice una prueba preliminar con mínimo un referente experto en el área específica (por ejemplo, para el caso de ejemplo de asincronías, un experto en medicina respiratoria) y evalúe junto con el profesional cuál es la técnica más apropiada para destacar el elemento que se desea clasificar.

En el caso de encuesta de clasificación de asincronías, para hacer clara la selección de la respiración que se desea clasificar, se evaluaron distintas opciones, tales como programar para que aparezcan líneas verticales en su inicio y su final o cambiar el color de las líneas de las señales. Sin embargo, se consultó con dos referentes en una prueba preliminar y en base a las perspectivas aportadas por estos profesionales de la salud, se decidió que la alternativa más clara fue cambiar el color de fondo de la respiración de interés.

En cuanto al cambio del color de fondo, se eligió un color gris más claro que el resto de las respiraciones anteriores y posteriores que tienen un color de fondo más oscuro. Además, se puede ver en la Figura 4 y Figura 5 que el cambio de color para destacar la respiración seleccionada no fue en detrimento de la distinción entre inspiración y espiración. Se puede ver que, tanto en la respiración seleccionada como en las adyacentes, la inspiración es de un color más claro que la espiración, dado que se considera que es útil separar visualmente de este modo ambas fases de la respiración para facilitar la clasificación.

Zoom

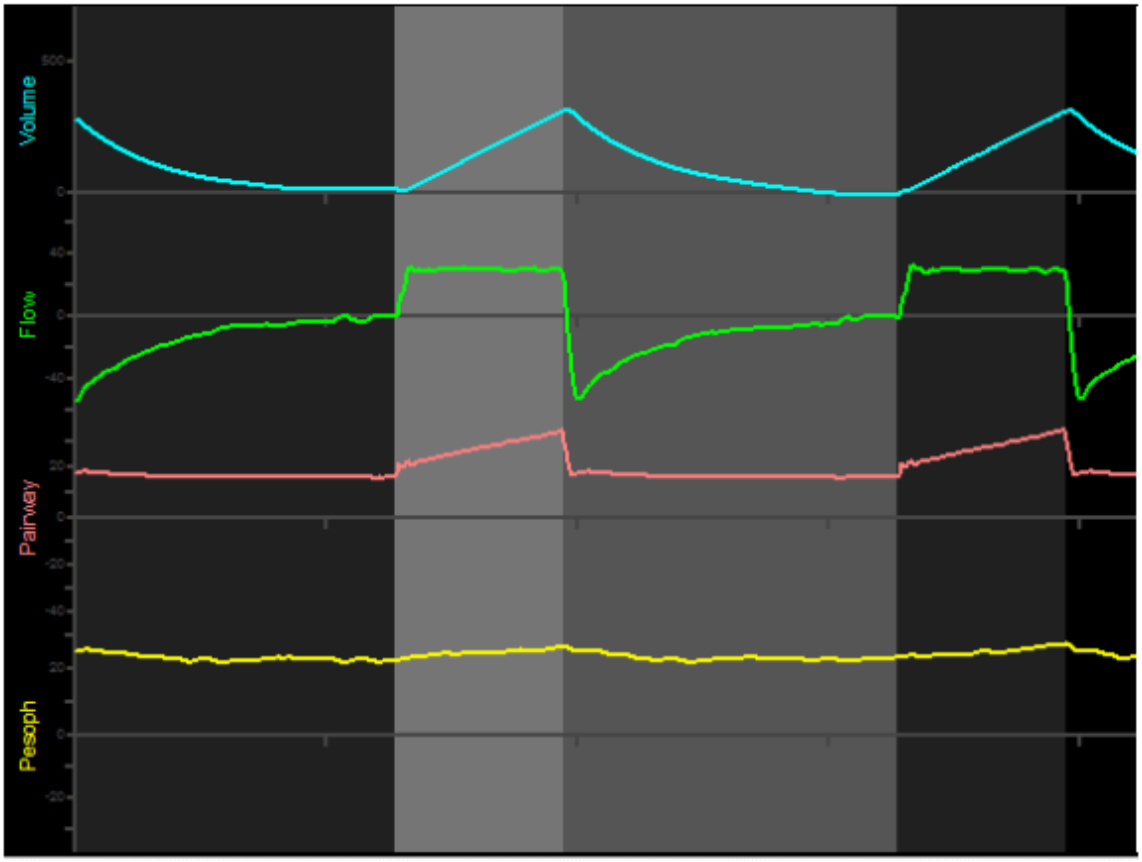


Figura 4. Imagen del patrón respiratorio del paciente incluyendo señales de volumen, flujo, presión de la vía aérea y presión esofágica, con zoom en la respiración de interés. Se puede ver la fase de inspiración en un gris más claro y la fase de espiración en un gris más oscuro.

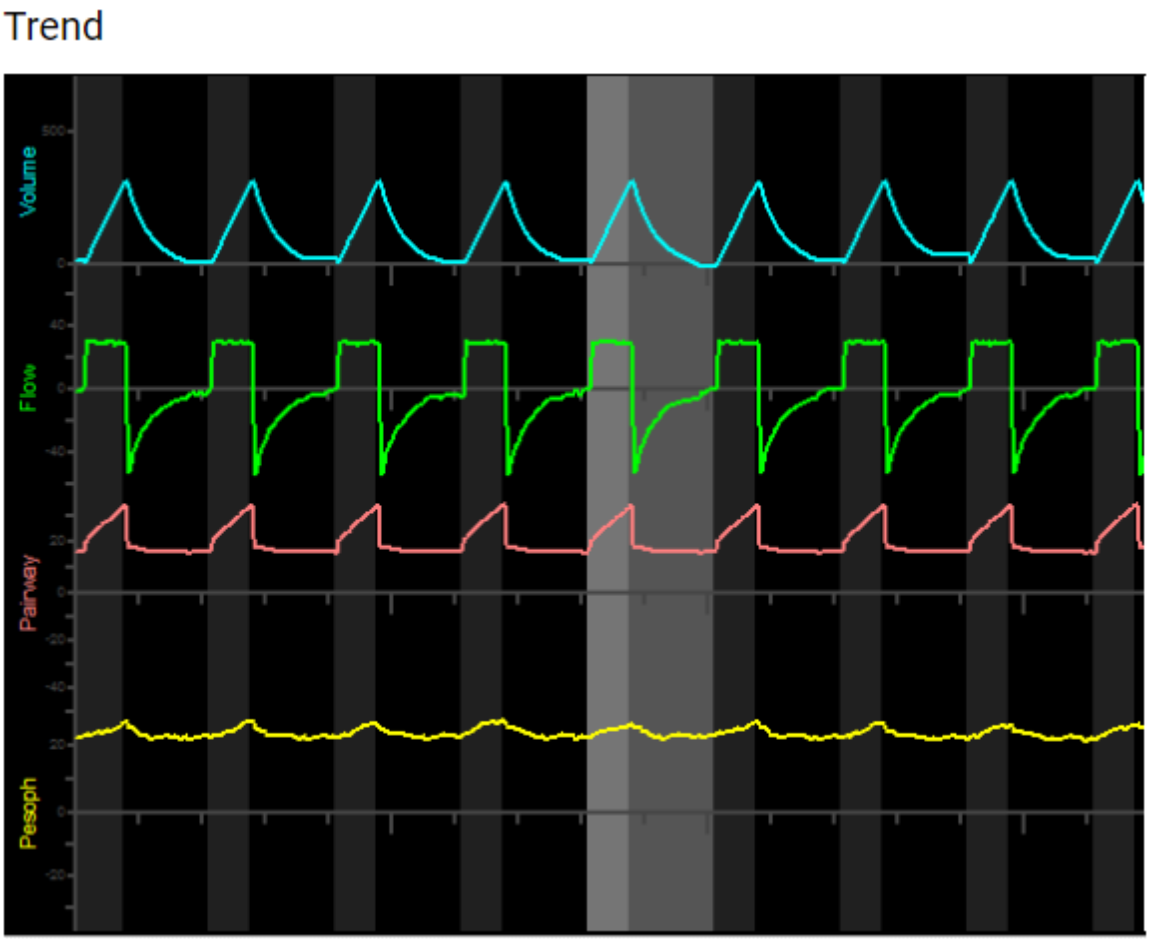


Figura 5. Imagen del patrón respiratorio del paciente, mostrando la tendencia en las respiraciones anteriores y posteriores a la respiración de interés, la cual se encuentra destacada en gris claro y es la misma mostrada en la Figura 4. Se distingue la fase de inspiración de la de espiración, empleando un tono de fondo más claro para la primera.

Cantidad de elementos a clasificar

Cada usuario que desee elaborar un Gold Standard debe definir la cantidad de elementos a clasificar que determinarán la cantidad de imágenes. En esta decisión influencian múltiples factores como el objetivo para el cual se quiere elaborar un Gold Standard, cuántos expertos participarán en el estudio y cuánto tiempo se estima que le toma la clasificación a un experto.

Una vez definida la cantidad de elementos a clasificar, se recomienda hacer una prueba preliminar complementaria previa al armado de la encuesta final que se enviará a todos los expertos en la cual

se evalúe la duración de la encuesta con esta cantidad de elementos elegida para ver si es apropiada o se debe modificar.

Para la encuesta realizada de clasificación de asincronías se decidió establecer una relación de compromiso entre la cantidad de calificaciones y el tiempo disponible y de concentración del médico experto. En este sentido, se decidió utilizar 100 respiraciones a clasificar por cada evaluador, es decir, se incluyeron 100 preguntas en la encuesta (un total de 200 imágenes, dos por cada respiración). Se tomó dicha decisión a partir de una prueba preliminar con dos referentes que demoraron entre treinta y cuarenta minutos en completar las 100 clasificaciones.

Determinación de las categorías

El usuario debe definir categorías para armar el Gold Standard acorde a la señal fisiológica y la patología que se quiera analizar en cada caso. Las mismas deben ser independientes y mutuamente excluyentes, para permitir posteriormente la realización de un análisis de acuerdo entre los evaluadores.

Para la encuesta de clasificación de señales respiratorias en asincronías paciente-respirador, se definieron, como ya fue mencionado, categorías para la clasificación de asincronías que son independientes y mutuamente excluyentes. Se consideraron aquellas asincronías que se encuentran asociadas en algunos casos como categorías en sí mismas, las cuales se caracterizan por la conjunción de dos asincronías dado que, justamente, dicha asociación define características particulares distintas a aquellas de las asincronías separadas.

Se decidió definir las categorías de las asincronías en inglés debido a que algunos de los expertos evaluadores residen en el exterior. Asimismo, dado que la literatura científica del tema se encuentra predominantemente en inglés, el idioma no fue una traba para los demás evaluadores ya que se encuentran familiarizados con los términos. Por último, dado que en varios casos una misma categoría de asincronía en inglés cuenta con diversas posibles traducciones al español, es posible que los expertos no utilicen los mismos términos en español para referirse a la misma asincronía y en inglés hay una menor ambigüedad.

Las categorías que se utilizaron son: “Ineffective effort (without reverse triggering)”, “Auto-triggering”, “Short cycling”, “Double triggering (without reverse triggering)”, “Short cycling + Double triggering”, “Long cycling”, “Reverse triggering + Ineffective efforts”, “Reverse triggering + Double triggering”, “No asynchrony is present” y “I don’t know”.

Se debe destacar que las categorías “Ineffective effort (without reverse triggering)”, “Auto-triggering”, “Short cycling”, “Double triggering (without reverse triggering)”, “Short cycling + Double triggering”, “Long cycling”, “Reverse triggering + Ineffective efforts” y “Reverse triggering + Double triggering” indican presencia de asincronías paciente-respirador. La única categoría que muestra ausencia de asincronías es “No asynchrony is present”. La categoría restante “I don’t know” implica que el clasificador no sabe con certeza a que categoría pertenece la respiración analizada.

Si se hubiera buscado hacer una encuesta con categorías dicotómicas, las mismas podrían haber sido “Hay presencia de asincronías” y “Hay ausencia de asincronías”. Para este caso de ejemplo, se decidió incluir a todas las clasificaciones de asincronías posibles para poder elaborar un análisis estadístico de acuerdo más exhaustivo (cuando hay solo dos categorías la coincidencia por azar es más probable) y porque se consideró que el impacto clínico del Gold Standard elaborado sería mayor de este modo, dado que se puede distinguir entre asincronías más y menos graves.

Modo de presentación

En cuanto al modo de seleccionar las categorías por parte de los evaluadores, en las encuestas con múltiples opciones existen dos posibilidades. Una permite que el clasificador seleccione todas las posibilidades que considere correctas y la segunda consiste en la elección de una única opción, por lo que el evaluador debe, para cada elemento a clasificar, elegir la opción que considere más idónea frente a otras que pudieran parecerle más o menos adecuadas.

Para la construcción de Gold Standards, dado que las categorías son mutuamente excluyentes, se debe utilizar un modo de mostrar las opciones en el que el evaluador pueda seleccionar una única opción y en ningún caso pueda elegir más de una. Dentro de las posibilidades de Google Forms, se eligió el formato *Mutiple Choice*.

A modo de ejemplo, en la Figura 6, se puede ver una imagen con la presentación final de las opciones de la encuesta de clasificación de respiraciones en asincronías que surge de las categorías elegidas junto con el formato *Mutiple Choice* de Google Forms.

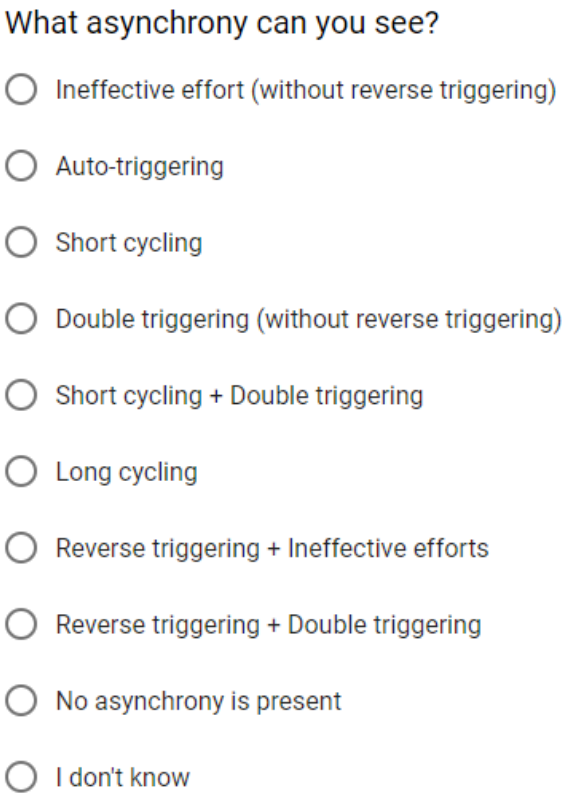


Figura 6. Presentación final de las opciones de la encuesta para el caso de ejemplo de clasificación de respiraciones en asincronías paciente-ventilador. Se pueden observar las categorías elegidas en el formato Mutiple Choice de Google Forms.

Otro aspecto relacionado al modo de presentación de la información es el orden en el que se muestra la información al evaluador en cada sección de la encuesta, es decir, para cada elemento a clasificar. Una opción es presentar primero la imagen del elemento a clasificar y, luego, las posibles categorías. Otra opción es mostrar primero las categorías y, después, la imagen del elemento a clasificar.

En el caso ejemplificador de la encuesta de clasificación de asincronías, de acuerdo con las opiniones de referentes recabadas en la prueba preliminar, se decidió mostrar primero las dos imágenes que representan la respiración (es decir, la representación gráfica del elemento a clasificar) y luego las

opciones de clasificación de asincronías. Considerando que el grupo de posibles categorías se mantiene constante en toda la encuesta (en otras palabras, se repite para cada elemento) y luego de varias clasificaciones sucesivas es posible que el evaluador comience a acordárselas o mínimamente estar familiarizado con estas, se buscó que la atención del experto se centrara en las imágenes de la respiración. Dichas imágenes cambian para cada sección ya que se trata de una nueva respiración cada vez.

Origen de las señales

Para cada Gold Standard se debe disponer de un repositorio de señales fisiológicas que se desea clasificar, a partir del cual se puede armar la encuesta de clasificación para expertos con la herramienta desarrollada. Este conjunto de señales debe ser provisto por el usuario para cada Gold Standard. Se recomienda que las señales fisiológicas a clasificar sean provenientes de distintos hospitales para minimizar el sesgo que podría provenir del centro de salud.

La herramienta desarrollada permite la clasificación de una misma base de señales en diferentes grupos de categorías si se utilizan las mismas señales para generar diferentes encuestas con distintas clasificaciones. Cada una de estas encuestas permite facilitar la generación de un Gold Standard distinto para las mismas señales.

En el caso de la encuesta de clasificación de asincronías, las señales fueron facilitadas por el Dr. Pablo Rodríguez quien lidera una investigación centrada en asincronías durante los primeros días de ventilación mecánica de pacientes con SDRA. El grupo poblacional particular del cual provienen las señales utilizadas en este estudio es de pacientes con SDRA que se encuentran en Unidad de Cuidados Intensivos con ventilación mecánica durante las primeras 72 horas de VM. Estas señales fueron registradas con equipos FluxMed.

Dado que las señales utilizadas en este estudio provienen de una investigación focalizada en detectar asincronías en pacientes con SDRA, se puede esperar que las asincronías más prevalentes en este grupo poblacional de alto riesgo sean aquellas asincronías como *reverse triggering* (asociada a peores resultados en la ventilación mecánica).

No obstante, no puede afirmarse que *reverse triggering* es una de las patologías más frecuentes de asincronías paciente-ventilador, dado que la frecuencia de una patología u otra depende de la población analizada. En futuras investigaciones en las que esta plataforma se utilice para señales respiratorias y para detectar asincronías, se pueden evaluar otros grupos de pacientes. Se espera que para cada población analizada, varíe el porcentaje de asincronías, la prevalencia de un cierto tipo de asincronía y también la proporción de distribución en distintas categorías de asincronías.

Es importante destacar que, para este caso de ejemplo, los registros provienen de cinco centros de salud distintos. Como ya fue mencionado, al utilizar señales provenientes de distintos hospitales se minimiza el sesgo por centro. Los hospitales involucrados son Hospital Italiano de Buenos Aires, Hospital Mitre, Hospital Churruca, Hospital Anchorena y CEMIC.

En la encuesta realizada, se muestran un total 20 respiraciones por hospital, que pertenecen a dos grupos de diez respiraciones sucesivas. Si bien hay respiraciones que son adyacentes, se las ordenó de forma aleatoria previamente a armar la encuesta de modo que los evaluadores no analizaron dichas respiraciones de forma sucesiva.

Generación de las imágenes

Cada usuario que quiera elaborar un Gold Standard debe contar con un software que permita generar imágenes a partir de la base de señales que se quiera clasificar. Cada señal fisiológica es distinta y se mide de forma diferente. Por eso, es importante disponer de un software que pueda representarla gráficamente, adaptándose a su unidad y escala.

Para el caso de ejemplo de señales respiratorias, se realizó en el marco de este trabajo una modificación del software FluxView. Se cambió el código fuente del software FluxView, diseñado originalmente para adquirir, registrar y leer datos de archivos de señales respiratorias, para programar nuevas funciones que permitan generar imágenes de respiraciones a partir de archivos de señales. Estas funciones que se agregaron al software pueden tomarse a modo de ejemplo de las funcionalidades que debe incluir el software que utilice el usuario para obtener imágenes de otras señales fisiológicas (no necesariamente respiratorias) que quiera clasificar cuando utilice la herramienta para obtener otros Gold Standards.

La modificación del software FluxView se realizó de modo que este pueda ser utilizado por un usuario en el proceso de elaboración de cualquier Gold Standard de señales respiratorias en general. En particular, las imágenes que se utilizaron en la encuesta de clasificación de señales respiratorias en asincronías paciente-respirador fueron generadas a partir de dicha modificación del FluxView, pero también podrían generarse imágenes para clasificación de modos ventilatorios, entre otras posibles clasificaciones de señales respiratorias.

Se decidió llamar FVC2 (el número 2 indica que es la segunda versión de la modificación) al software FluxView modificado. Las funcionalidades adicionales respecto al software FluxView original son: seleccionar una respiración de modo que se cambie el color de fondo de esta, tomar una captura de pantalla de la imagen de la respiración mostrada y tomar capturas sucesivas de las respiraciones posteriores registrando el número de respiración, las cuales serán almacenadas en una misma carpeta. Estas funciones nuevas son las que debería permitir cualquier software que decida utilizar un usuario para obtener imágenes de las señales fisiológicas que quiera clasificar: seleccionar el elemento a clasificar, tomar una captura de las curvas de las señales y almacenarlas en una carpeta tomando un registro del número de elemento.

En las modificaciones del programa se tuvo en cuenta que el mismo sea rápido y fácil de utilizar para un usuario no experto en informática. Una vez abierto un archivo .flux (aquel archivo que almacena las señales) y modificando los parámetros manualmente de acuerdo a cada señal, el usuario no demora más de cinco minutos en obtener imágenes de todas las respiraciones de la señal. De hecho, una vez ajustada la escala de tiempo y las amplitudes de las diferentes señales manualmente y seleccionada la primera respiración, el usuario puede obtener capturas de pantalla de todas las respiraciones de forma automática mediante la única acción de seleccionar el botón de “Captura de pantalla” de la barra de herramientas.

Todas las imágenes generadas a partir de un mismo archivo de señales se almacenan en una carpeta denominada “NombreArchivoElegido_FVC2_XX”, siendo NombreArchivoElegido el nombre del archivo de señales seleccionado y XX el número de ejecución de dicho software para la misma señal. Dentro de cada carpeta, el nombre de los archivos que son imágenes de respiraciones es “NombreArchivoElegido_FVC2_XX_YY”, siendo YY el número de respiración.

Como protocolo, que se explica posteriormente en la sección “Instrucciones a seguir por el usuario que desea armar una plataforma”, el usuario inicialmente ajusta la escala de tiempo para hacer un zoom en una única respiración y ejecuta el programa por primera vez. Luego, debe reajustar la escala de tiempo para que se vean aproximadamente entre 9 y 15 respiraciones en pantalla y ejecutar el programa nuevamente.

Se consideró la opción de que el software ajuste la escala de amplitud de las señales de forma automática, tomando como referencia el punto máximo y mínimo identificado en el archivo para cada señal en particular. No obstante, se descartó esta alternativa ya que en varios archivos de señales existen problemas de *overflow* en la medición, esporádicos, sin relevancia fisiológica. Por lo tanto, el usuario debe ajustar la escala de cada señal individualmente, ya que la magnitud de la señal varía según cada paciente (y sus propias características fisiológicas, como *compliance* pulmonar, si posee alguna patología como EPOC o SDRA, entre otros factores) y luego descartar aquellas respiraciones con *overflow*.

En el programa, como se ha mencionado, se incluyó la opción de ajustar tanto la escala de tiempo, en el eje horizontal, como la amplitud de las señales (independientemente para cada señal mostrada) en el eje vertical. También, se pueden seleccionar cuáles señales respiratorias se desean mostrar según el Gold Standard que se quiera generar. En particular, para este estudio se utilizan el volumen, el flujo, la presión de la vía aérea y la presión esofágica, pero en caso de que el usuario desee elaborar un Gold Standard diferente para señales respiratorias, el mismo podría seleccionar más señales (por ejemplo, la curva de capnografía, la presión gástrica, la presión transdiafragmática y la impedancia eléctrica) o quitarse algunas de estas que no sean relevantes para el caso de señales nomencladas que quiera construir.

Se ilustrará en una sucesión de varias imágenes el funcionamiento del software FluxView modificado para generar imágenes de respiraciones para el caso de ejemplo de clasificación de asincronías. Su funcionamiento será similar si se utiliza para generar imágenes para generar un Gold Standard de señales respiratorias distinto, aunque puede variar la selección de señales.

En la Figura 7 se puede ver una imagen del software FVC2 en funcionamiento apenas se carga una señal. En la barra de herramientas, se puede ver el ícono  que es el que se utiliza para realizar la

captura de pantalla sucesiva de todas las respiraciones registrando el número de respiración. Todas estas imágenes se almacenan en una misma carpeta. Dentro de las opciones de señales que el software puede mostrar a partir de un archivo de señales, en esta imagen se puede ver que las señales seleccionadas son el volumen (en celeste), el flujo (en verde), la presión de la vía aérea (en color rojo) y la presión esofágica (en color amarillo).

En la Figura 8 se puede ver la misma señal que en la Figura 7 una vez que el usuario ha ajustado la amplitud de las señales (debe destacarse que la amplitud puede modificarse para cada señal mostrada independientemente) y ha hecho zoom en el eje del tiempo en la primera respiración.

Por su lado, en la Figura 9 se puede visualizar al aspecto de la primera respiración una vez que ha sido seleccionada: el color de su fondo es gris más claro en comparación con el resto de las respiraciones

Por último, en la Figura 10 se muestra un ejemplo de imagen que genera el FluxView y que se almacena en la carpeta, registrando el número de respiración.

Figura 7. Software FluxViewC2 utilizado para tomar captura de pantalla de respiraciones. Se puede ver el ícono  *que es el que se debe utilizar para tomar captura de pantalla sucesiva de todas las respiraciones. Dentro de las opciones de señales que el software puede mostrar a partir de un archivo Flux, se muestra el volumen, el flujo, la presión de la vía aérea y la presión esofágica.*

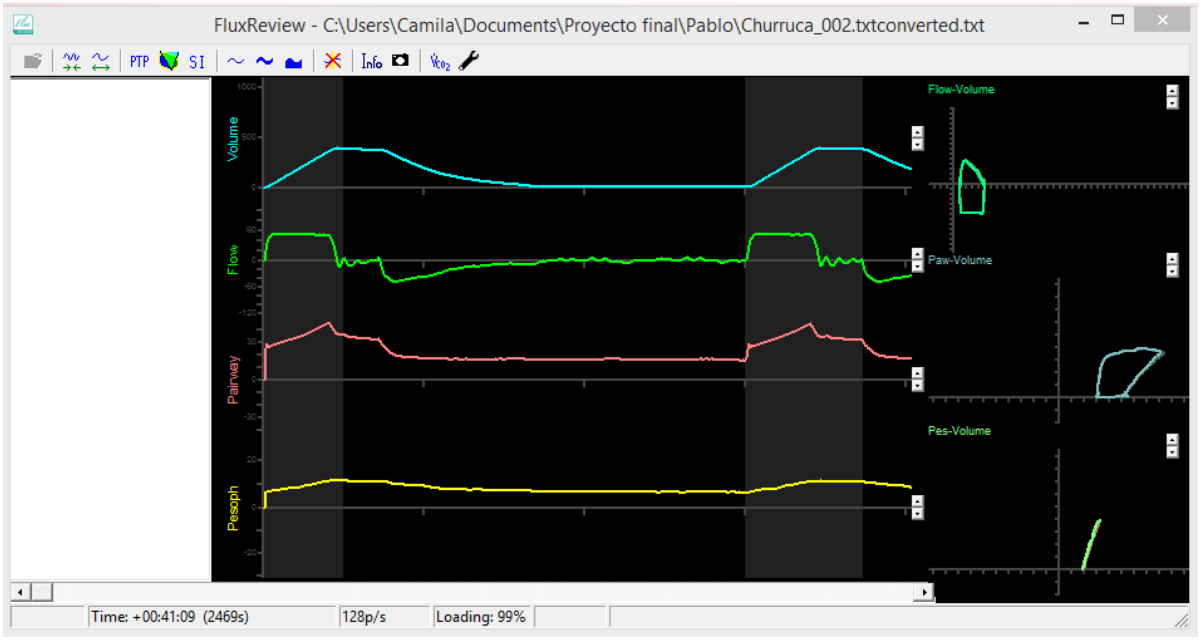


Figura 8. Software FluxViewC2 utilizado para tomar captura de pantalla de respiraciones. Se han ajustado la amplitud de las señales, independientemente entre sí utilizando las flechas en la barra ubicada a la derecha de cada una de las señales. También se han utilizado los íconos de la barra de herramientas superior para ajustar la escala de tiempo (horizontal) hasta hacer zoom en la primera respiración.

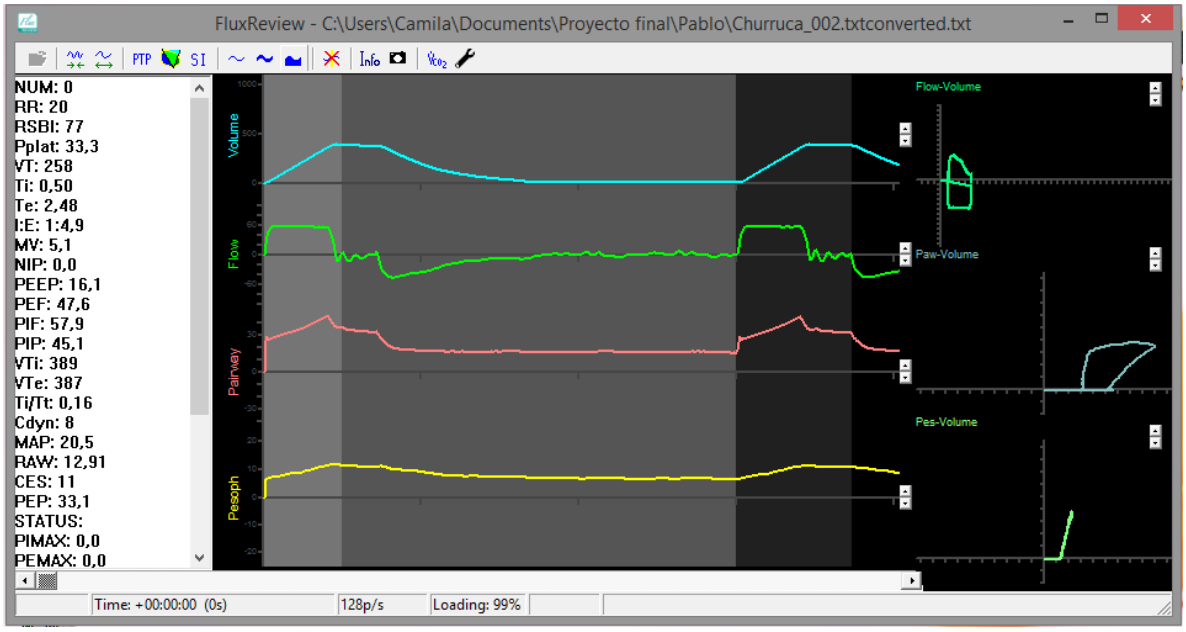


Figura 9. Software FluxViewC2 utilizado para tomar captura de pantalla de respiraciones. Se ha seleccionado manualmente la primera respiración de la misma señal que en la Figura 8. Se pueden visualizar en la columna de la izquierda varios parámetros que describen a la respiración, pero que no son relevantes para esta aplicación en particular, por lo que no serán considerados para generar la imagen.

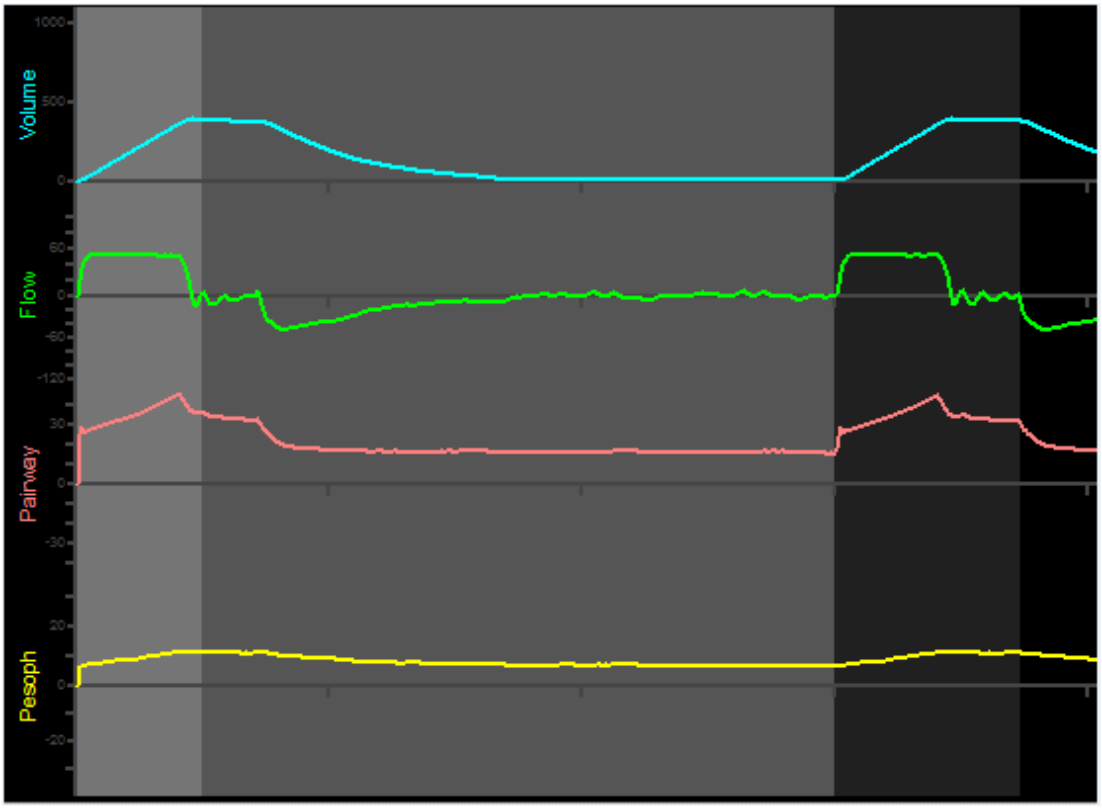


Figura 10. Se puede ver una de las imágenes generadas por el FluxView para la señal, en particular es la imagen generada correspondiente a la primera respiración de la señal mostrada en la Figura 9.

Evaluadores

Para generar un Gold Standard es imprescindible contar con profesionales de la salud que sean expertos en el tema de interés para que cumplan la función de evaluadores. De acuerdo con el tipo de clasificación que se desee hacer, se puede requerir un especialista con más o menos experiencia. En general, es de interés para elaborar un Gold Standard contar con la colaboración de especialistas en el área específica que se trate con la mayor experiencia profesional y trayectoria posible. El usuario de la plataforma debe decidir qué requisitos necesita que cumplan los profesionales de la salud que serán los clasificadores de su encuesta para cada Gold Standard que quiera desarrollar.

Para el caso de clasificación de respiraciones en asincronías paciente-ventilador, la experiencia del especialista en el área de detección de asincronías es necesaria para una buena sensibilidad en la detección, lo cual es un requisito esencial si la plataforma de encuestas debe servir para generar un Gold Standard [18].

En un estudio realizado por Ramirez et al destinado a evaluar la habilidad de los profesionales de la salud que trabajan en UCI para detectar asincronías haciendo un análisis de formas de onda [19], se encontraron diferencias estadísticamente significativas cuando se compararon los resultados de los profesionales de la salud con y sin entrenamiento previo en ventilación mecánica según el número de asincronías detectadas correctamente. Por ejemplo, respecto a los profesionales que identificaron 3 asincronías, 63 (81%) fueron médicos entrenados mientras que 15 (19%) eran no entrenados; para el caso de 2 asincronías, 72 (65%) estaban entrenados en comparación con 39 (35%) no entrenados; para detectar 1 asincronía, 55 (47%) eran entrenados y 61 (53%), no entrenados; para el caso de 0 asincronías, hubo 17 (28%) entrenados y 44 (72%) no entrenados. Se ilustran dichas proporciones en la Figura 11. Adicionalmente, en dicho estudio, la profesión y la experiencia trabajando en UCI no demostraron ser un factor de relevancia.

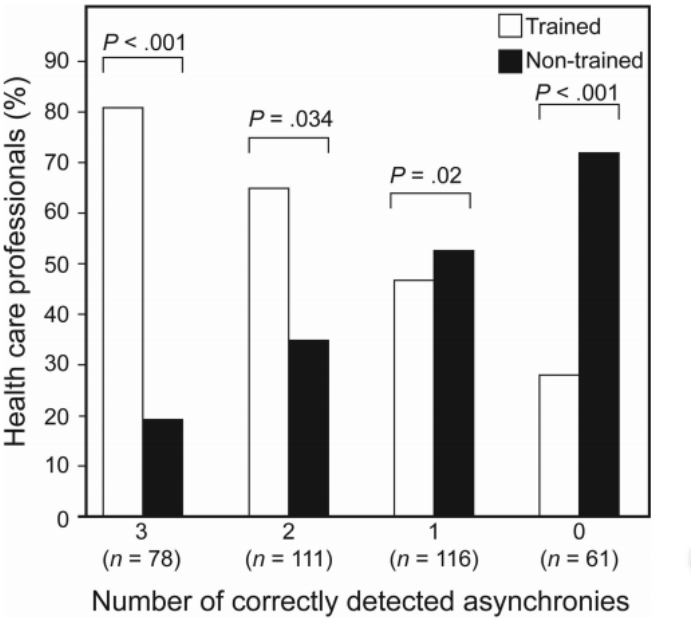


Figura 11. Comparación intragrupo entre profesionales de la salud entrenados y no entrenados para la detección de asincronías, según el número de asincronías reconocidas correctamente. El entrenamiento es un factor relevante para la detección de asincronías. La imagen fue obtenida de la publicación "Ability of ICU Health-Care Professionals to Identify Patient-Ventilator Asynchrony Using Waveform Analysis" de Ramirez et al [19].

En el caso de ejemplo de clasificación de asincronías paciente-respirador, se decidió incluir como evaluadores a profesionales de la salud especialistas en medicina respiratoria. Se incluyó tanto a médicos como a kinesiólogos, todos los cuales cuentan con entrenamiento y experiencia en detección de asincronías.

En total, participaron siete (7) expertos como evaluadores, por lo que cada una de las respiraciones tendrá siete clasificaciones debido al diseño totalmente cruzado. Se trata de 4 médicos y 3 kinesiólogos, todos los cuales son profesores en varias Sociedades de Terapia Intensiva, cuentan con publicaciones en el área de asincronías paciente-ventilador y su desempeño profesional en Terapia Intensiva es mayor a 10 años.

Nótese que en este trabajo los términos evaluadores, observadores, codificadores y expertos se utilizan como equivalentes e intercambiables entre sí sin implicar una diferencia en significado. Cualquiera de esos términos puede ser utilizado para nombrar a los profesionales de la salud que clasifican las respiraciones.

Instrucciones a seguir por el usuario que desea armar una plataforma

El usuario que desee generar un Gold Standard para señales respiratorias debe:

Respecto al programa FVC2 (FluxViewC versión 2):

- Abrir el archivo de señales que se quiere clasificar.
- Elegir cuáles de las señales registradas desea mostrar en la imagen(por ejemplo, volumen, presión de la vía aérea, flujo, presión esofágica, presión gástrica).
- Ajustar la escala de amplitud de las señales que se decidió mostrar del archivo elegido (no se puede determinar por default ya que cambian las amplitudes según las características fisiológicas del paciente medido). La amplitud de cada una de las señales mostradas debe modificarse independientemente con las flechas.
- Marcar con el *mouse* la primera respiración, de modo que quede seleccionada (queda marcada con un color más claro)

- Ajustar la escala de tiempo (horizontal), de modo que la imagen quede centrada en una sola respiración.
- Seleccionar el botón de la cámara (captura de pantalla) en la barra de herramientas.
- Esperar mientras se toman las capturas de pantalla con zoom en una respiración, a lo largo de toda la señal.
- Todas las imágenes con Zoom se guardan en una carpeta denominada NombreArchivoElegido_FVC2_00, que se puede localizar donde el usuario decida. Se registra para cada imagen el número de respiración.
- Volver a ejecutar el FVC2. y abrir el archivo
- Ajustar la escala de tiempo de modo que queden en pantalla aproximadamente 10 respiraciones (pueden ser más o menos según el criterio del usuario), con el objetivo de ver la tendencia de paciente. Ajustar la escala vertical si es necesario.
- Marcar la primera respiración, de modo que quede seleccionada.
- Seleccionar el botón de "Captura de pantalla" en la barra de herramientas.
- Esperar mientras se toman las capturas de pantalla de toda la señal.
- Todas las imágenes que muestran tendencia se guardan en una carpeta denominada NombreArchivoElegido_FVC2_01. El usuario debe decidir dónde se ubica dicha carpeta. Se registra para cada imagen el número de respiración, al igual que con las imágenes con zoom en una respiración.
- Repetir el proceso para todos los archivos que se desean incluir en el Gold Standard.
- Copiar todas las carpetas dentro de la carpeta Imágenes en Google Drive.

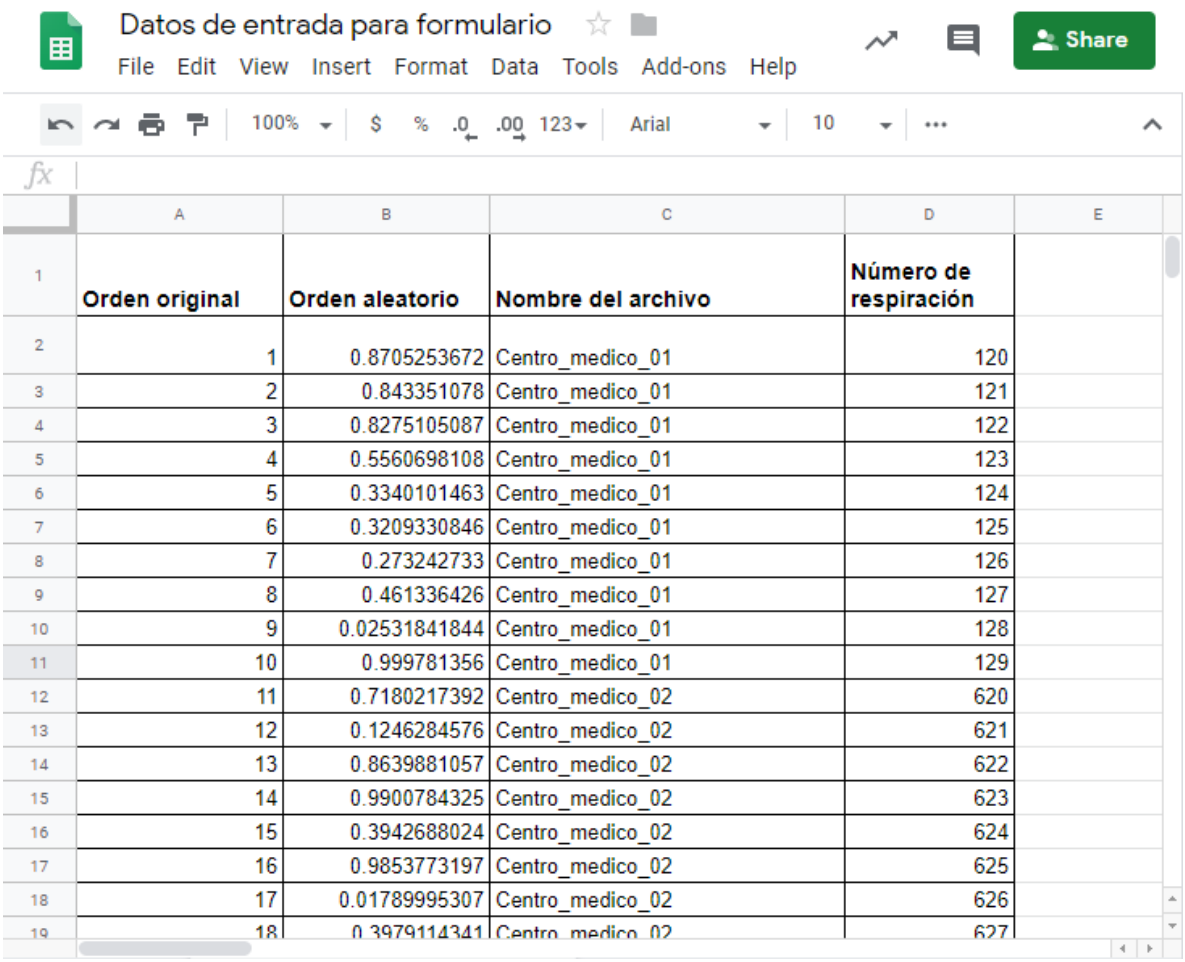
En el *Spreadsheet* "Datos de entrada para el formulario":

- Escribir en la columna "Nombre de archivo" (C), el nombre del archivo que desea, y en la columna (D) llamada "Número de respiración", la respiración deseada correspondiente dicho archivo. Se generan tantas filas como respiraciones seleccione. Puede visualizarse el

aspecto del *Spreadsheet* utilizado para crear la encuesta de clasificación de respiraciones en asincronías en la Figura 12.

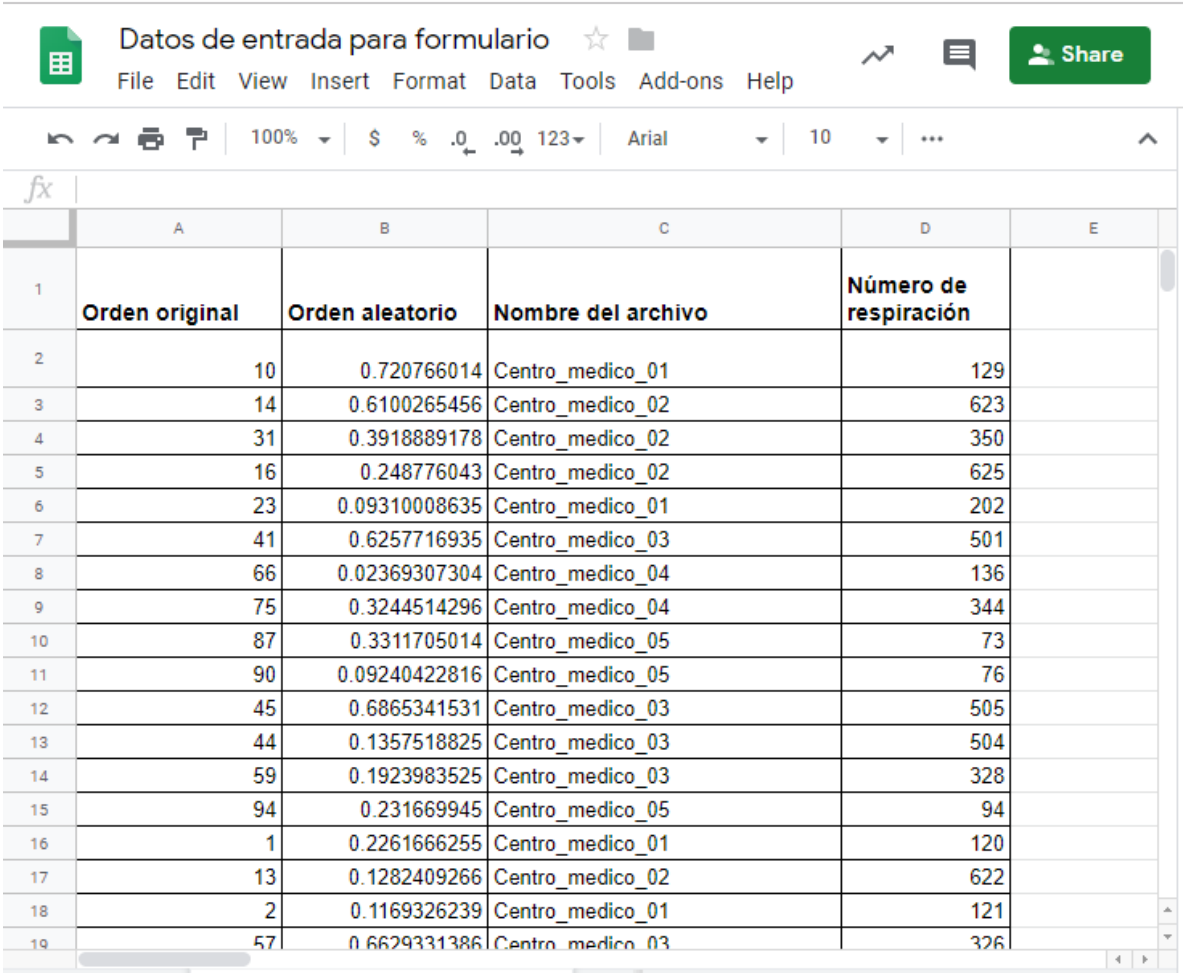
- La columna B, llamada “Orden aleatorio”, siempre que hayan datos en la columna C (es decir haya un nombre de archivo), se autocompletará con un número aleatorio entre 0 y 1, determinado por la función RAND(). Esto permitirá ordenar las respiraciones para que se le presenten en un orden al azar al experto evaluador y reducir su sesgo.
- El usuario, cuando termine de escribir todos los archivos con sus correspondientes respiraciones, debe ordenar de acuerdo a la columna B (en la que se encuentran números aleatorios) de forma Z-->A. Se puede ver un ejemplo de cómo quedan ordenadas al azar las respiraciones seleccionadas en la Figura 13.

Por último, el usuario debe ejecutar el Script “Armar Form”, que crea una encuesta a partir de la información definida por el usuario en el *Spreadsheet*. En el *Spreadsheet* se almacena la información del orden en el que se le presentan las opciones a los especialistas que sirve para realizar el Gold Standard en última instancia (permite saber a qué respiración corresponde determinada clasificación).



	A	B	C	D	E
1	**Orden original**	**Orden aleatorio**	**Nombre del archivo**	**Número de respiración**	
2	1	0.8705253672	Centro_medico_01	120	
3	2	0.843351078	Centro_medico_01	121	
4	3	0.8275105087	Centro_medico_01	122	
5	4	0.5560698108	Centro_medico_01	123	
6	5	0.3340101463	Centro_medico_01	124	
7	6	0.3209330846	Centro_medico_01	125	
8	7	0.273242733	Centro_medico_01	126	
9	8	0.461336426	Centro_medico_01	127	
10	9	0.02531841844	Centro_medico_01	128	
11	10	0.999781356	Centro_medico_01	129	
12	11	0.7180217392	Centro_medico_02	620	
13	12	0.1246284576	Centro_medico_02	621	
14	13	0.8639881057	Centro_medico_02	622	
15	14	0.9900784325	Centro_medico_02	623	
16	15	0.3942688024	Centro_medico_02	624	
17	16	0.9853773197	Centro_medico_02	625	
18	17	0.01789995307	Centro_medico_02	626	
19	18	0.3979114341	Centro_medico_02	627	

Figura 12. Imagen del Spreadsheet utilizado para armar la encuesta de clasificación de respiraciones en asincronías. El usuario debe completar el nombre del archivo del cual se desean extraer las imágenes y los números de los elementos de interés (en este caso, número de respiración de interés). La columna de orden aleatorio se autocompleta a medida que el usuario completa los campos "Nombre de archivo" y "Número de respiración".



Datos de entrada para formulario

File Edit View Insert Format Data Tools Add-ons Help

Share

100% $ % .0 .00 123 Arial 10

	A	B	C	D	E
1	**Orden original**	**Orden aleatorio**	**Nombre del archivo**	**Número de respiración**	
2	10	0.720766014	Centro_medico_01	129	
3	14	0.6100265456	Centro_medico_02	623	
4	31	0.3918889178	Centro_medico_02	350	
5	16	0.248776043	Centro_medico_02	625	
6	23	0.09310008635	Centro_medico_01	202	
7	41	0.6257716935	Centro_medico_03	501	
8	66	0.02369307304	Centro_medico_04	136	
9	75	0.3244514296	Centro_medico_04	344	
10	87	0.3311705014	Centro_medico_05	73	
11	90	0.09240422816	Centro_medico_05	76	
12	45	0.6865341531	Centro_medico_03	505	
13	44	0.1357518825	Centro_medico_03	504	
14	59	0.1923983525	Centro_medico_03	328	
15	94	0.231669945	Centro_medico_05	94	
16	1	0.2261666255	Centro_medico_01	120	
17	13	0.1282409266	Centro_medico_02	622	
18	2	0.1169326239	Centro_medico_01	121	
19	57	0.6629331386	Centro_medico_03	326	

Figura 13. Imagen del Spreadsheet utilizado para armar la encuesta de clasificación de respiraciones en asincronías, una vez que el usuario ordena las filas en función de la columna "Orden aleatorio". El formulario mostrará los elementos de la manera en la que aparecen en este formulario. Esta información es útil para el proceso de reconstrucción de Gold Standard, ya que permite conocer a qué respiración y señal de origen corresponde determinada clasificación.

Análisis estadístico

Tipo de variables

La herramienta desarrollada para simplificar el proceso de creación de Gold Standards genera como resultado en primera instancia un conjunto de datos, obtenidos de las respuestas de clasificación de los expertos evaluadores. El usuario de la herramienta utilizará estos datos para generar la base de señales nomecladas, es decir, el Gold Standard. El procesamiento estadístico que realice de los datos dependerá del objetivo que busque el usuario en cada caso. Es relevante el tipo de datos con el que se cuenta dado que determinará cuáles métodos de estadística se utilizarán para analizarlos. Los datos, en general, y las variables utilizadas en encuestas, en particular, se pueden clasificar como variables nominales, ordinales, de intervalo o de relación.

Las variables nominales (categóricas) son nombres de categorías o clases desordenadas, sin relación jerárquica entre ellas. Por su lado, las variables ordinales son conjuntos de categorías en el que el orden adquiere importancia (por ejemplo, estadificación de un cáncer), si bien la magnitud o número específico asignado a una cierta categoría no tiene importanica (por lo cual tampoco tienen sentido la mayoría de operaciones aritméticas sobre estas variables). Por su parte, las variables de intervalo refieren a un conjunto ordenado de categorías con el requisito adicional de que las categorías formen una serie de intervalos que sean exactamente de la misma magnitud (por ejemplo, la diferencia entre una temperatura de 37 ° C y 38 ° C es de 1 grado, que es igual a la que existe entre 38 ° C y 39 ° C). Sin embargo, en una escala de intervalo no existe un punto cero absoluto que indique una ausencia completa de la variable que se está midiendo, por lo que las relaciones de valores no son significativas. En cambio, una variable de relación tiene las características de una variable de intervalo pero agrega un punto cero absoluto, por lo cual proporciones de números en la escala reflejan proporciones de magnitud para la variable que se mide [27],[28].

En este caso, al igual que en el resto de estudios realizados para generar Gold Standards, los resultados que se obtienen son en base a categorías nominales. Algunas de las variables nominales, denominadas variables dicotómicas, tienen solo 2 valores posibles (por ejemplo: si una característica específica está presente o ausente); por otro lado, otras variables nominales pueden

tener varios valores posibles. El número real de categorías puede ser determinado por el investigador de acuerdo al Gold Standard a obtener.

Para el caso de ejemplo de clasificación de asincronías, se podrían haber establecido dos categorías dicotómicas (presencia y ausencia de asincronía) o bien varias opciones de categorías posibles. Se elige la segunda opción y se establecen una clasificación que tiene diez posibles categorías de asincronías. Asimismo, para poder realizar un análisis más exhaustivo se utilizan categorías mutuamente excluyentes.

Tipo de encuesta

En cuanto al diseño de la encuesta, se utiliza en el ejemplo de uso de la herramienta para el caso de clasificación de respiraciones en asincronías un diseño totalmente cruzado con múltiples evaluadores (múltiples significa más de dos). La cantidad de clasificaciones por respiración es constante y es realizada por los mismos evaluadores, no por evaluadores aleatorios seleccionados de un conjunto más grande y distintos para cada respiración.

Fiabilidad inter-evaluadores e intra-evaluadores

Los datos son el fundamento esencial para el razonamiento y el cálculo. Para que las respuestas de los expertos sean útiles, las mismas deben ser generadas con todas las precauciones posibles respecto a los contaminantes conocidos, distorsiones y sesgos, intencionales o accidentales y, además, deben significar lo mismo para todos aquellos que las utilicen. Esto es particularmente relevante para este Proyecto en el cual los datos obtenidos por la herramienta desarrollada tienen el fin de ser utilizados para generar un Gold Standard.

Existen dos concepciones respecto a la fiabilidad, de las cuales se desprendan maneras de realizar operaciones. Por un lado, existe concepción de la teoría de medición de la fiabilidad que considera que la fiabilidad se basa en la seguridad de que proporciona que los datos se obtienen independientemente de la persona que evalúa o instrumento utilizado. En consecuencia, un procedimiento de clasificación es confiable cuando responde a los mismos fenómenos de la misma manera, independientemente de variaciones en el proceso. Por otro lado, existe otra concepción que considera que la fiabilidad es el grado en que los miembros de una comunidad designada están

de acuerdo en las interpretaciones de datos dados. Esta es una concepción interpretativa de la fiabilidad [29], que es la utilizada para el proceso de generación de Gold Standards que se busca facilitar en este trabajo.

Para el análisis elaborado para el caso de ejemplo de clasificación de respiraciones en asincronías, se utilizarán medidas de fiabilidad entre evaluadores (en inglés, Inter-Rater Reliability que se abrevia con las siglas IRR). Se trata de una medición de la medida en que los evaluadores asignan la misma categoría (para este estudio, el mismo tipo de asincronía) a la misma variable (la misma respiración). Su importancia radica en el hecho de que representa la medida en que los datos recopilados son representaciones correctas de las variables medidas.

La fiabilidad entre evaluadores es un tema de interés para le elaboración de Gold Standards debido al hecho de que varias personas pueden interpretar los fenómenos de interés de manera diferente. En particular, como fue señalado, para realizar una clasificación de asincronías, se requiere experiencia y entrenamiento en el área [20] e incluso entre expertos no hay un completo consenso, por lo que la consistencia entre las respuestas de los evaluadores cobra especial relevancia [30].

En realidad, existen dos categorías de fiabilidad con respecto a los recopiladores de datos: la fiabilidad en múltiples recolectores de datos, que es la ya mencionada fiabilidad inter-evaluadores, y por otro lado la fiabilidad de un solo recolector de datos, que se denomina fiabilidad intra-evaluador. En este último caso, con un solo recolector de datos se evalúa si en el caso en el que se le presenten exactamente la misma respiración (mismas imágenes), el experto interpreta las curvas de la misma manera y clasifica la misma respiración. En otras palabras, se evalúa si el experto es consistente consigo mismo para realizar las clasificaciones. La herramienta de encuestas desarrollada puede servir para esta aplicación también, pero no se realizará en el presente trabajo ya que en esta instancia el objetivo es utilizarla de modo que facilite generar un Gold Standard (es decir, el foco de su utilización es evaluar tipos de asincronías en respiraciones), no para evaluar al clasificador en sí mismo.

Existen diferentes tipos de conceptos relacionados a la fiabilidad: estabilidad, reproducibilidad y exactitud. Su distinción no se basa en cómo se mide el acuerdo, sino en base a cómo los datos de fiabilidad son obtenidos. En primer lugar, la estabilidad se puede obtener a partir de un diseño de

prueba-repetición (de la prueba) y la causa del desacuerdo son las inconsistencias intra-observador. Se mide en la medida en que un procedimiento de codificación produce los mismos resultados en ensayos repetidos. Un observador vuelve a obtener la misma información y recategoriza, generalmente después de un cierto tiempo. Las inconsistencias del observador consigo mismo son un factor inevitable que no puede eliminarse de ningún estudio en el que participe algún observador o calificador. La falta de fiabilidad se manifiesta en las variaciones en el rendimiento de un observador, también llamadas desacuerdo intra-observador o inconsistencias individuales, pueden deberse a la inseguridad, el descuido, las distracciones o la tendencia a relajar los estándares de rendimiento frente al cansancio. La estabilidad, la forma más débil de fiabilidad, es insuficiente como único criterio para aceptar datos como confiables.

En cuanto a la reproducibilidad, el diseño del experimento puede definirse como de prueba-prueba (es decir, implica que se realicen pruebas a más de un observador); por ejemplo, en este dos o más individuos, que trabajan de forma independiente, clasifican los mismos objetos (las mismas respiraciones) en base a un mismo conjunto categorías. La reproducibilidad es el grado en que un proceso puede ser replicado por diferentes evaluadores que trabajan bajo diferentes condiciones, en diferentes ubicaciones. Las causas del desacuerdo son, ya no sólo las inconsistencias intra-observador, sino también los desacuerdos entre observadores. En este estudio, estos desacuerdos en la interpretación y la codificación de las respiraciones pueden originarse en el hecho que todos los evaluadores son profesionales de la salud con conocimiento de medicina respiratoria, pero tienen distintas profesiones (tanto médicos como kinesiólogos) y formaciones académicas y, además, se desempeñan en sus funciones en diferentes contextos clínicos. En comparación con la estabilidad, la reproducibilidad es una medida mucho más sólida de fiabilidad [29].

Por último, la exactitud implica un diseño prueba-estándar y las causas del desacuerdo son las inconsistencias intra-observador, desacuerdos entre observadores y, fundamentalmente, las desviaciones respecto a un estándar. Esto no es alcanzable para este trabajo dado que, precisamente, el objetivo es construir una herramienta que permita obtener datos precisos para obtener un estándar en áreas en las que el mismo no existe [29].

La información relativa a los distintos tipos de fiabilidad está resumida en la Tabla 3.

Fiabilidad	Diseño de pruebas	Causas del desacuerdo	Fuerza
Estabilidad	Prueba-repetición	Inconsistencias intra-observador	Débil
Reproducibilidad	Prueba-prueba	Inconsistencias intra-observador + desacuerdos entre observadores	Mediana
Exactitud	Test-estándar	Inconsistencias intra-observador + desacuerdos entre observadores + desvíos de un estándar	Fuerte

Tabla 3. Tipos de conceptos de fiabilidad, su respectivo diseño de pruebas, causas del desacuerdo y su fuerza. Esta tabla fue tomada del libro de Klaus Krippendroff llamado "Content analysis : an introduction to its methodology" [29].

Para este caso de ejemplo de clasificación de asincronías, al igual que en cualquier otro caso de generación de Gold Standards con la herramienta desarrollada, se analiza en base al concepto de reproducibilidad, que es la medida de fiabilidad más posible que se puede alcanzar sin disponer de un estándar aceptado (de hecho, el estándar es justamente lo que se busca construir). Dado que la prueba involucró a varios expertos interviene tanto el factor de inconsistencia intra-observador como los desacuerdos entre observadores.

Datos para el análisis estadístico

En estudios en los que proporcionar calificaciones es costoso o requiere mucho tiempo, la selección de un subconjunto de sujetos para el análisis de fiabilidad entre evaluadores puede ser más práctica porque requiere menos calificaciones generales, y la fiabilidad entre evaluadores para el

subconjunto de sujetos se puede usar para generalizar a la muestra completa de sujetos involucrados en la clasificación [31].

Sin embargo, en este caso de ejemplo de clasificación de señales respiratorias en asincronías, se decidió utilizar todos los datos y respuestas disponibles para analizar la fiabilidad entre evaluadores, es decir, se consideraron las 100 respiraciones clasificadas y los 7 (siete) evaluadores involucrados para los análisis.

Descripción vs. Modelado

Se debe distinguir entre describir y modelar el nivel de acuerdo. Dentro de las maneras de describir el acuerdo, se pueden mencionar la proporción de veces que concuerdan dos clasificaciones del mismo caso, la proporción de veces que los evaluadores concuerdan con categorías específicas, la cantidad de veces que diferentes evaluadores utilizan los diversos niveles de calificación, etc.

La cuantificación del acuerdo de cualquier otra manera implica un modelo sobre cómo se hacen las calificaciones y por qué los evaluadores están de acuerdo o en desacuerdo. Este modelo es explícito, como con los modelos de estructura latente, o implícito, como los coeficientes kappa [32].

Descripción estadística

Acuerdo porcentual (percent agreement)

Si bien ha habido una variedad de métodos para medir la fiabilidad entre evaluadores, tradicionalmente se midió como porcentaje de acuerdo (es decir, de coincidencia entre respuestas), calculado como el número de respuestas iguales dividido por el número total de respuestas.

Entre sus principales ventajas, se deben mencionar que es una estadística de fácil cálculo e interpretación. Sin embargo, el acuerdo porcentual no tiene en cuenta la posibilidad de que los evaluadores coincidan por azar y, por lo tanto, puede sobreestimar el verdadero acuerdo. Para profundizar en el acuerdo por conjeturas de los expertos se hicieron análisis posteriores con la estadística kappa, que al ser comparados con el acuerdo porcentual permiten analizar si existe una sobreestimación del acuerdo. En el caso de ejemplo de clasificación de asincronías respiratorias, dado que los evaluadores son especialistas en medicina respiratoria y están entrenados en

clasificación de asincronías, es probable que existan pocas conjeturas y, en este contexto, el acuerdo porcentual *a priori* es una medida descriptiva adecuada.

$$Percent\ agreement = \frac{Número\ de\ respuestas\ coincidentes}{Número\ de\ respuestas\ totales}\ x\ 100$$

Cuando hay dos evaluadores, para obtener la medida del porcentaje de acuerdo, se construye una matriz en la que las columnas representan a los diferentes evaluadores y las filas, a las variables (o viceversa). Las celdas en la matriz contienen los datos para cada variable que los evaluadores ingresaron. El porcentaje de acuerdo se calcula de forma independiente para cada elemento que haya sido objeto de clasificación y luego el porcentaje de acuerdo general se computa promediando los porcentajes de acuerdo correspondientes a cada uno de los elementos.

En el caso de múltiples evaluadores, el porcentaje de acuerdo de cada elemento se calcula considerando el acuerdo entre todos los posibles pares de evaluadores. Luego, el porcentaje de acuerdo general se calcula promediando todos los acuerdos porcentuales de los elementos individuales, al igual que en el caso de dos evaluadores.

En el caso de ejemplo de clasificación de asincronías, dado que hay más de dos evaluadores, se implementó en Google Scripts un algoritmo que obtiene programáticamente las respuestas de los expertos del *Spreadsheet*, calcula el porcentaje de acuerdo para cada respiración considerando cada par posible de evaluadores (21 en total en este estudio con 7 evaluadores) y luego promedia todos los porcentajes.

El acuerdo porcentual total teniendo en cuenta todos los evaluadores y las 100 respiraciones para este trabajo fue de 81,9%. Este valor indica que hay una gran concordancia considerando el total de 100 respiraciones. Además, la mediana de los valores de acuerdo porcentual para las 100 respiraciones es de 100%. Esto es consistente con un alto nivel de acuerdo entre los evaluadores.

Si se calcula el acuerdo porcentual considerando únicamente las 89 respiraciones en las que hay un acuerdo de 5, 6 o 7 evaluadores (representan un porcentaje superior al 71,4% de todos los evaluadores para esta encuesta), es decir, excluyendo las 11 respiraciones con bajo nivel de consenso, con el fin de obtener un Gold Standard de mayor precisión, el porcentaje de acuerdo es de 88%. La mediana de los 89 acuerdos porcentuales correspondientes a las 89 respiraciones

también es 100%, al igual que en el caso de las 100 respiraciones. La mediana fue igual y se obtiene un porcentaje de acuerdo superior que indica una concordancia que es superior respecto al caso anterior que computa los valores para las 100 respiraciones totales. Estos eran resultados esperables teniendo en cuenta que se excluyen las 11 respiraciones con bajo nivel de acuerdo.

Existe la posibilidad de construir un Gold Standard sólo con las 89 respiraciones de mayor nivel de acuerdo si se quiere tener un Gold Standard más preciso (debido a los niveles más altos de consenso), pero de todas maneras debe destacarse que el nivel de acuerdo fue alto en ambos casos, para las 100 respiraciones totales y para las 89 respiraciones con mayor nivel de consenso. La elección depende del usuario y sus objetivos al elaborar la base de datos nomenclada, en última instancia.

Cantidad de respiraciones con diferentes niveles de acuerdo

Complementariamente al acuerdo porcentual, se busca conocer cuántas de las respiraciones analizadas tienen un determinado nivel de acuerdo. El acuerdo porcentual es una medida adecuada para cuantificar el acuerdo general y sintentizarlo en un solo número de fácil interpretación, pero es necesario hacer un análisis adicional para distinguir la cantidad de respiraciones que cuentan con diferentes niveles de acuerdo. Esto se debe a que no es lo mismo contar con 50 respiraciones con 100% de acuerdo entre los evaluadores y otras 50 respiraciones con un 30% de acuerdo, que contar con 100 respiraciones con un 75% de acuerdo.

Con el fin de describir la distribución de respiraciones en los niveles de acuerdo, se elabora una tabla (Tabla 4) en la que se cuentan la cantidad de respiraciones con todos los posibles niveles de acuerdo

Se debe señalar que el porcentaje máximo de evaluadores de acuerdo no es igual al acuerdo porcentual (las únicas excepciones son los casos en el que estén todos los evaluadores de acuerdo y en el que ninguno lo está, para los cuales ambos valores serían 100% y 0% respectivamente). El porcentaje máximo de evaluadores de acuerdo se calcula tomando el máximo número de evaluadores de acuerdo y mostrando porcentualmente cuánto representa dicha cantidad (la opinión mayoritaria) respecto del total de evaluadores (en este caso, el total es de 7 evaluadores). Por otro lado, el acuerdo porcentual tiene en cuenta todos los posibles acuerdos entre pares. Por ejemplo, para el caso de cuatro evaluadores de acuerdo en una determinada categoría, el

porcentaje máximo de evaluadores de acuerdo es de 57,1%. Sin embargo, el acuerdo porcentual para esa respiración puede ser de 42,9% si los tres evaluadores restantes están de acuerdo en otra categoría, 33,3% si dos de los evaluadores restantes están de acuerdo en otra categoría y uno está en desacuerdo con todos los demás, y 28,6% si están los tres evaluadores restantes están todos en descuerdo con los primeros cuatro y entre sí.

Cantidad de evaluadores de acuerdo	Porcentaje de evaluadores de acuerdo	Porcentaje de respiraciones
7	100,0 %	56 %
6	85,7 %	27 %
5	71,4%	6 %
4	57,1 %	6 %
3	42,9 %	5 %
2	28,6 %	0 %
1	14,3 %	0 %
0	0,0%	0 %

Tabla 4. Porcentaje de respiraciones para cada determinado nivel de acuerdo, expresado en cantidad máxima de evaluadores de acuerdo y el porcentaje que estos representan respecto del total de calificadores.

Existe un porcentaje de evaluadores de acuerdo superior o igual al 71,4% (es decir, al menos 5 de 7 evaluadores eligen la misma categoría) para 89 respiraciones, que representan el 89% del total de respiraciones analizadas. Incluso, para 83% del total de respiraciones, hay un porcentaje de 85,7% o 100,0% de calificadores de acuerdo (en otras palabras, 6 o 7 evaluadores de acuerdo respectivamente). Asimismo, no existe ninguna respiración (0% del total) en la que la cantidad de evaluadores de acuerdo sea inferior a 3 (42,9% del total) e incluso las respiraciones con tres evaluadores de acuerdo como máximo (42,9% del total) representan solamente el 5% del total de las respiraciones analizadas.

El alto nivel de consenso entre evaluadores para este caso de ejemplo muestra la utilidad de los datos generados para permitir la construcción de un Gold Standard. Estos datos son resultado de la

herramienta desarrollada y, aun contando con la colaboración de los expertos, no habría sido posible obtenerlos de manera sencilla y rápida sin la misma. De esta manera, se puede ver el potencial de simplificación del proceso de creación de Gold Standards a partir de esta herramienta

Frecuencia de categorías por evaluador

Es útil conocer la cantidad de veces que cada evaluador utiliza cada categoría, para saber si un cierto clasificador tiene una mayor tendencia a utilizar una determinada categoría en comparación a los demás. Para ello, se programó una herramienta que registra a partir del *Spreadsheet* de respuestas de la plataforma, la cantidad de veces que cada evaluador utilizó cada categoría (Tabla 5). Además, se crea un gráfico de barras a partir de dichos datos para facilitar un análisis comparativo entre la cantidad de veces que los distintos evaluadores eligen cada categoría (Figura 14). En este, es sencillo visualizar la proporción entre la selección entre las distintas categorías, tanto para el mismo evaluador como entre los evaluadores entre sí.

Del total de clasificaciones de respiraciones, se puede ver que el 34,3% indica presencia de asincronías, mientras que el 59,4% corresponde a la ausencia de asincronía y el 6,3% muestra que el clasificador no supo clasificar la respiración.

Se puede observar que el 31,3% de las clasificaciones totales se encuadran en las categorías “Reverse triggering + Ineffective efforts” y “Reverse triggering + Double triggering”, mientras que las categorías restantes que indican presencia de asincronías representan solo el 3%. Tiene sentido que las categorías más prevalentes dentro de las asincronías son las que involucran *reverse triggering* teniendo en cuenta el origen de las señales utilizadas para este caso de ejemplo. Se trata de pacientes que se encuentran en la UCI y tienen síndrome SDRA, los cuales se caracterizan por tener asincronías como *reverse triggering*. Como ya fue mencionado, este resultado era esperable a priori, pero esto no implica que en otras poblaciones *reverse triggering* sea una de las asincronías más comunes. De hecho, para las respiraciones de cada grupo de pacientes es posible que exista una distribución distinta entre las categorías de asincronías.

La frecuencia con la que los clasificadores eligen una determinada categoría es similar entre sí. La diferencia entre la máxima y mínima cantidad de veces que cada categoría fue elegida fue de: 4 para “Ineffective Effort (without Reverse Triggering)”, 3 para “Auto-Triggering”, 1 para “Short cycling”, 2

para “Double Triggering (without Reverse Triggering)”, 0 para “Short cycling + Double triggering”, 4 para “Long cycling”, 7 para “Reverse Triggering + Ineffective efforts”, 9 para “Reverse triggering + Double triggering”, 21 para “No asynchrony is present” y 16 para “I don’t know”. Estos valores representan una diferencia porcentual de 4%, 3%, 1%, 2%, 0%, 4%, 7%, 9%, 21% y 16%, entre la máxima y mínima cantidad de veces que fue seleccionada cada categoría, respectivamente, teniendo en cuenta el total de clasificaciones por evaluador (100).

Puede considerarse, teniendo en cuenta la frecuencia similar con la que los expertos seleccionan una cierta categoría, que el sesgo por evaluador es mínimo. En otras palabras, no existe una tendencia muy marcada de un determinado evaluador a seleccionar una categoría, que permita distinguirlo de los demás evaluadores. Estos resultados respaldan las conclusiones de alto consenso alcanzadas en base al acuerdo porcentual y a la distribución de respiraciones por niveles de acuerdo.

Evaluador	**Ineffective effort (without reverse triggering)**	**Auto triggering**	**Short cycling**	**Double triggering (without reverse triggering)**	**Short cycling + Double triggering**	**Long cycling**	**Reverse triggering + Ineffective efforts**	**Reverse triggering + Double triggering**	**No asynchrony is present**	**I don’t know**
1	0	0	0	0	0	0	13	23	64	0
2	4	0	1	1	0	4	17	19	53	1
3	0	1	0	0	0	0	12	17	60	10
4	0	0	0	0	0	1	14	22	62	1
5	0	1	0	0	0	0	11	14	66	8
6	1	1	0	0	0	0	10	14	66	8
7	0	3	0	2	0	1	14	16	45	16
Totales parciales	5	6	1	3	0	5	77	109	416	44
% del total	0,7%	0,9%	0,1%	0,4%	0,0%	0,9%	13,0%	18,3%	59,4%	6,3%

Tabla 5. Frecuencia de selección de categorías según cada evaluador.

Se analizó el caso del clasificador 7 dado que, como se puede ver en la Figura 14, es aquel que selecciona la menor cantidad de veces la categoría “No asynchrony is present” y la mayor cantidad de veces “I don’t know”. En particular, utiliza la categoría “No asynchrony is present” 45 veces mientras que el promedio de los otros expertos (los clasificadores 1, 2, 3, 4, 5 y 6) es de 61,8 veces para la misma categoría. En cuanto a la categoría “I don’t know”, el evaluador 7 selecciona la misma 16 veces mientras que el promedio de los restantes seis clasificadores da una selección de 4,7 veces de esa categoría.

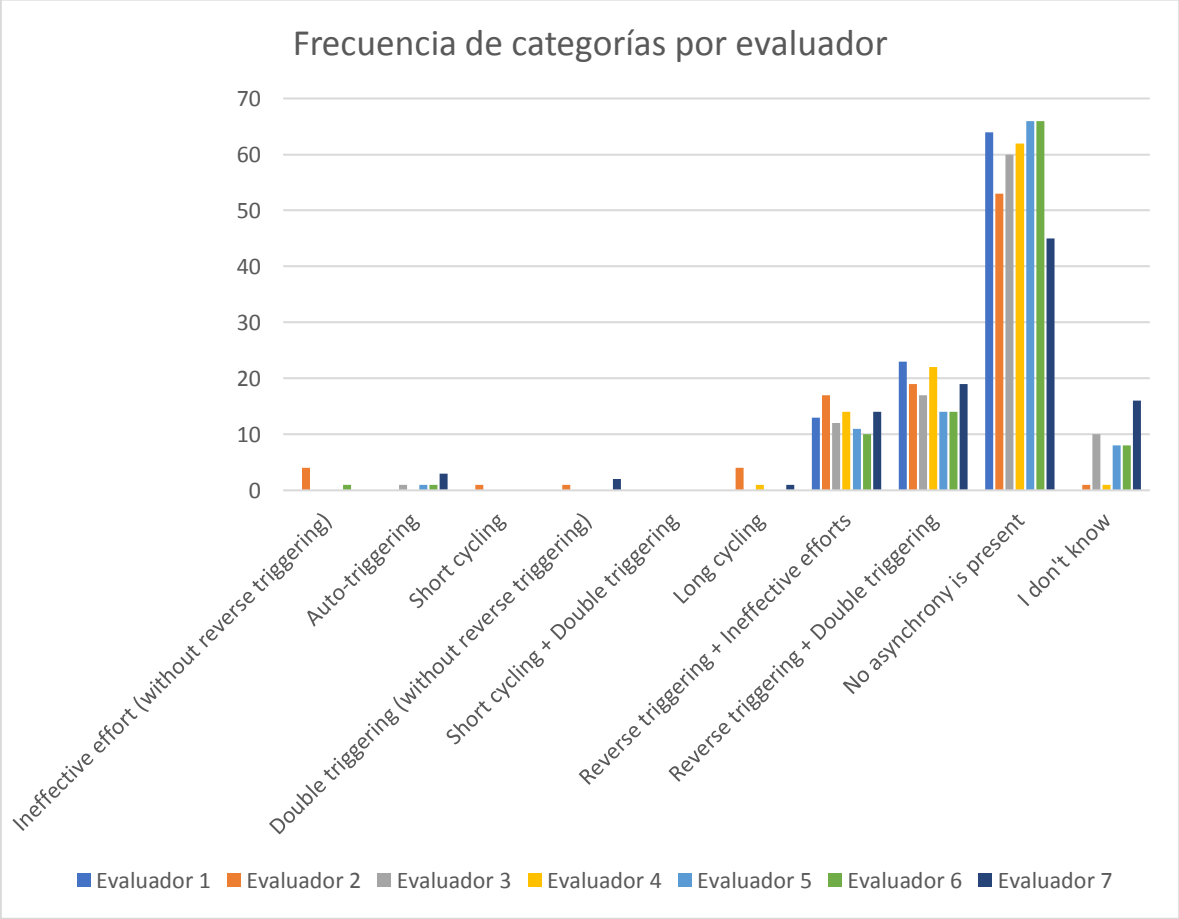


Figura 14. Representación visual de la frecuencia de selección de categorías según cada evaluador a partir de la información de la Tabla 5.

Por lo tanto, dado que a simple vista parecería ser que las evaluaciones del clasificador 7 son una fuente de desacuerdo para estas categorías en particular, se buscó identificar si hay respiraciones en las que el resto de los evaluadores seleccionen la categoría “No asynchrony is present” y el

evaluador 7 en su lugar elija “I don’t know”. Se encontraron 12 respiraciones (12% del total de las respiraciones analizadas) en las que los evaluadores 1, 2, 3, 4, 5 y 6 coincidieron en la elección de la categoría “No asynchrony is present” mientras que el evaluador 7 eligió “I don’t know”. Se puede considerar que el desacuerdo para los casos mencionados es relativo dado que el evaluador 7 no eligió una categoría de asincronías distinta al resto de los evaluadores, sino que consideró que no estaba seguro de si había una asincronía o no en la respiración en cuestión. Por lo tanto, si existen casos de desacuerdo en los que el mismo relativo, cualitativamente se podría considerar que el nivel de acuerdo podría ser incluso superior al indicado por las estadísticas descriptivas.

Distribución de errores entre los evaluadores

Se busca analizar si los errores son aleatorios y, por lo tanto, se distribuyen de manera similar entre todos los evaluadores y variables, o si un evaluador en particular registra con frecuencia valores diferentes de los demás.

Para este análisis, se tuvieron en cuenta únicamente las respiraciones en las que la cantidad de evaluadores de acuerdo es igual o mayor a 5, es decir, aquellas respiraciones en las que el consenso es mayor. Se distinguen estas respiraciones respecto de otras respiraciones en las que el máximo número de evaluadores de acuerdo es 1, 2, 3 y 4, las cuales se consideran como respiraciones con bajo consenso (en particular, algunas de ellas se analizarán cualitativamente en la sección “Análisis de respiraciones con bajo consenso”).

Por lo tanto, el número total de respiraciones consideradas para este análisis es de 89, en las cuales hay 39 veces en las que hay uno o más evaluadores en desacuerdo con la respuesta mayoritaria. En particular, 27 veces son correspondientes a casos en los que un único evaluador está en desacuerdo con los seis evaluadores restantes (es decir, se trata de aquellas 27 respiraciones con acuerdo de seis evaluadores) y 12 veces provienen de casos en los que dos evaluadores difieren de los cinco evaluadores restantes (es decir, se trata de las 6 respiraciones en las que hay acuerdo de cinco evaluadores).

La cantidad de veces que los evaluadores están en desacuerdo en respiraciones con alto nivel de consenso se pueden ver en la

Tabla *6* y son ilustrados en la Figura 15. Cabe destacar que, respecto al total de respiraciones clasificadas, las respiraciones con bajo consenso representan una proporción muy baja.

Evaluador	Desacuerdos
1	1 (1,1 %)
2	10 (11,2 %)
3	2 (2,3 %)
4	2 (2,3 %)
5	3 (3,4 %)
6	4 (4,5 %)
7	17 (19,1 %)

Tabla 6. Se expresa la cantidad de veces que los diferentes evaluadores están en desacuerdo con la respuesta mayoritaria y cuánto representa dicha cantidad del total de respiraciones consideradas (89). El análisis se hace para las 89 respiraciones en las que hay un acuerdo entre 5 o 6 clasificadores restantes.

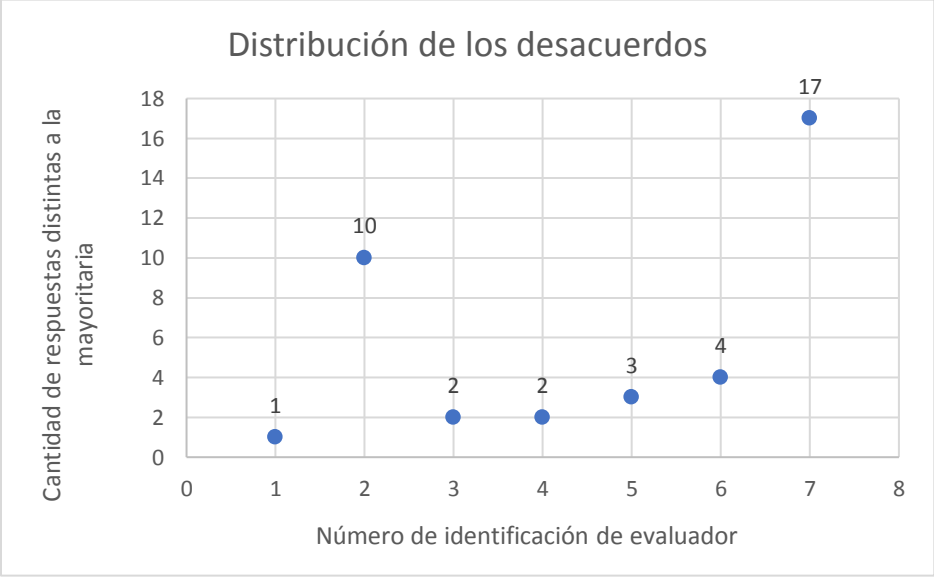


Figura 15. Se ilustra gráficamente la distribución de la cantidad de desacuerdos en base a la información de la Tabla 6.

Se puede ver en la Figura 15 que hay una cierta dispersión entre la distribución de errores. Se hizo un análisis para saber si la distribución de errores se asemejaba a la distribución normal y, como era de esperar observando que la mayoría de los valores se encuentran alrededor de 2, el resultado fue

negativo. Por lo tanto, algunas estadísticas como desvío estándar, no pueden ser utilizadas para describir esta variable. La cantidad de errores media es de 5,6 y la mediana es 3. Los dos evaluadores que registran una cantidad de errores superior a la media son el evaluador 2 y el evaluador 7. Se analizaron estos dos casos por separado para analizar si existe un cierto sesgo del evaluador.

El evaluador 2 tuvo 10 diferencias de opinión respecto a la mayoría. En el 90% de estas (9 de 10) en las que el evaluador difiere, la mayoría coincidió en la ausencia de asincronías (es decir, hay acuerdo en la categoría "No asynchrony is present"). Por su parte, el evaluador 2 consideró en 4 respiraciones que se trataba de "Ineffective effort (without reverse triggering)"; en otras 3, "Reverse triggering + Ineffective efforts" y en 2, "Long cycling". En el caso restante, que representa el 10% de las diferencias de opinión, los evaluadores coincidieron en la categoría "Reverse triggering + Double triggering" mientras que el evaluador consideró que se trata de un "Double triggering (without reverse triggering)". Este desacuerdo usualmente se origina en una diferencia de criterio para rotular respiraciones y no en un desacuerdo de proceso fisiológico. Los evaluadores están de acuerdo en el aspecto más esencial: hay un doble disparo (*double triggering*). Sin embargo, la diferencia radica en si se trata de una condición en que la respiración del doble disparo es activada por un esfuerzo muscular generado por el ventilador ("en sentido inverso"). Usualmente, el fenómeno de *reverse triggering* unido a *double triggering* involucra dos respiraciones relacionadas y los especialistas difieren en el modo de nomenclarlas.

Por lo tanto, para el caso del evaluador 2, dado que se puede presumir que la discrepancia en la mayoría de los casos radica en una diferencia de criterio para clasificar y no en el fenómeno fisiológico, se podría decir que se trata de un desacuerdo relativo y no puede considerarse que el este evaluador tenga mayor tendencia a tener errores que los demás expertos.

Por su parte, el evaluador 7 registra 17 opiniones divergentes respecto a la mayoría. En el 100% de ellas la categoría en la que hay consenso es "No asynchrony is present", es decir, la mayoría de los evaluadores coinciden en la ausencia de asincronías. En 13 de las 17 respiraciones (representan 76,5% de los desacuerdos en cuestión) el evaluador 7 no selecciona una categoría de asincronías, sino que elige la categoría "I don't know". Si bien esto no puede considerarse como acuerdo total, se trata de un desacuerdo relativo dado que el evaluador no difiere realmente de los demás calificadores, sino que sencillamente no sabe con certeza si hay asincronías o no. Si no se tuvieran

en cuenta como reales “desacuerdos” aquellas respiraciones en las que la mayoría de los evaluadores coinciden en que no hay asincronías, mientras que el evaluador en cuestión responde que no sabe, este tendría solamente 4 divergencias respecto de la opinión mayoritaria. En 3 de ellas selecciona la categoría, “Auto-triggering” y en una de ellas “Reverse triggering + Ineffective efforts”. Para este evaluador en particular, se debe tener en cuenta que se educó y se desempeña como profesional de salud en otro país, por lo cual su formación académica y contexto clínico puede ser un factor de influencia en sus opiniones.

Para el caso del evaluador 7, considerando los desacuerdos relativos, al igual que en el caso del evaluador 2, no puede considerarse que tenga mayor tendencia a cometer errores de clasificación respecto a los demás expertos y sus calificaciones son tenidas en cuenta para todos los análisis estadísticos realizados.

Análisis de respiraciones con bajo consenso

Una vez realizado un análisis cuantitativo de cuánto diverge cada evaluador respecto a la mayoría en las respiraciones, se hizo un análisis cualitativo de las respiraciones con bajo nivel de acuerdo con el fin de detectar causas posibles para el desacuerdo entre los expertos. Se consideraron aquellas respiraciones con un consenso máximo de 3 evaluadores (que representan 42,9% del total de evaluadores).

Debe destacarse que, si bien dichas respiraciones representan solamente el 5% del total de respiraciones clasificadas totales, es interesante analizarlas dado que, si se pueden detectar causas del desacuerdo, se podrían encontrar métodos para aumentar el consenso. Se analizó cada de estas cinco respiraciones individualmente.

Respiración con bajo consenso N1

La primera respiración con bajo consenso que se analizó se puede ver en la Figura 16 y los resultados de las clasificaciones en la Tabla 7. Se trata de la respiración 367 de uno de los archivos provenientes del Centro médico 4.

En las respuestas, se puede ver que, excluyendo la respuesta “I don’t know” del evaluador 2, el 50% de los evaluadores restantes consideraron que la respiración está asociada al *Double triggering* (dos

evaluadores eligieron la categoría “Reverse triggering + Double triggering” y uno, “Double triggering (without reverse triggering)”) y el otro 50% consideró que ninguna asincronía está presente en esa respiración (“No asynchrony is present”).

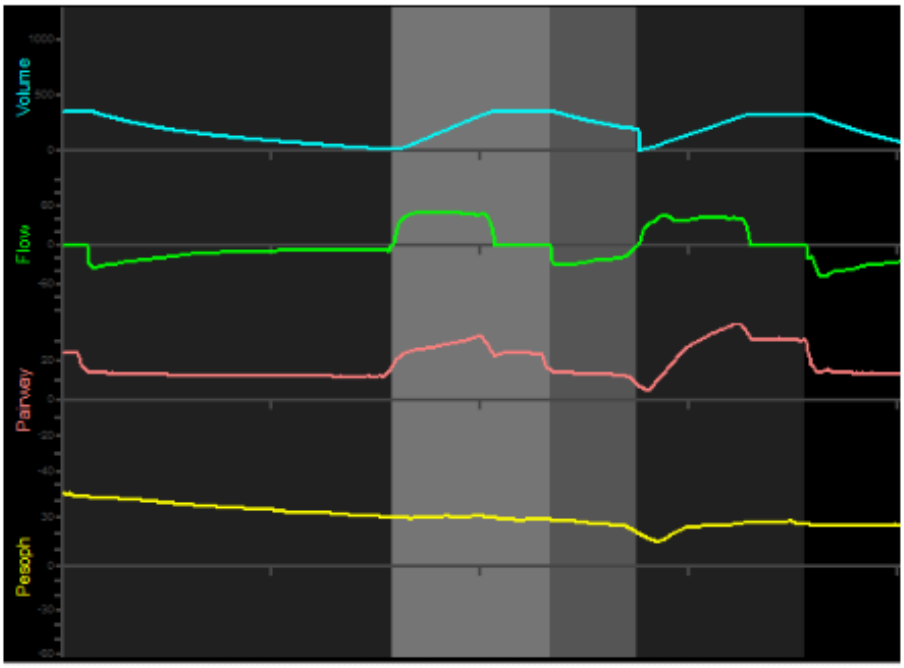
Evaluador	Respuesta
1	Reverse triggering + Double triggering
2	I don’t know
3	No asynchrony is present
4	Reverse triggering + Double triggering
5	No asynchrony is present
6	No asynchrony is present
7	Double triggering (without reverse triggering)

Tabla 7. Resultados de las evaluaciones de expertos para la respiración 367 del Centro médico 4.

Esta respiración en particular muestra un caso habitual de desacuerdo y debate entre expertos en el área de asincronías respiratorias. Si bien hay 3 evaluadores que consideran que no hay asincronía y uno no sabe clasificar la respiración, puede presumirse que el desacuerdo no radica en el fenómeno respiratorio que tiene lugar, sino en la forma de clasificar a las dos respiraciones asociadas que involucra este fenómeno.

La respiración marcada es mandatoria, es decir, es disparada por el respirador. Sin embargo, el gas que se administra para insuflar al paciente provoca que haga un esfuerzo (se observar en la Figura 16 en la deflexión negativa de la Peso) que, a su vez, dispara al respirador nuevamente y, de este modo, genera otra respiración. Por lo tanto, el proceso fisiológico en su conjunto consiste en el respirador que dispara al paciente que a su vez vuelve a disparar al respirador. Este fenómeno ya fue explicado y se conoce como *reverse triggering*. En esta situación, tanto la respiración marcada en la Figura 16 como la siguiente están asociadas y forman parte del mismo fenómeno de la interacción paciente-respirador.

Zoom (the highlighted breath is the same as the one below)



Trend (the highlighted breath is the same as the one above)

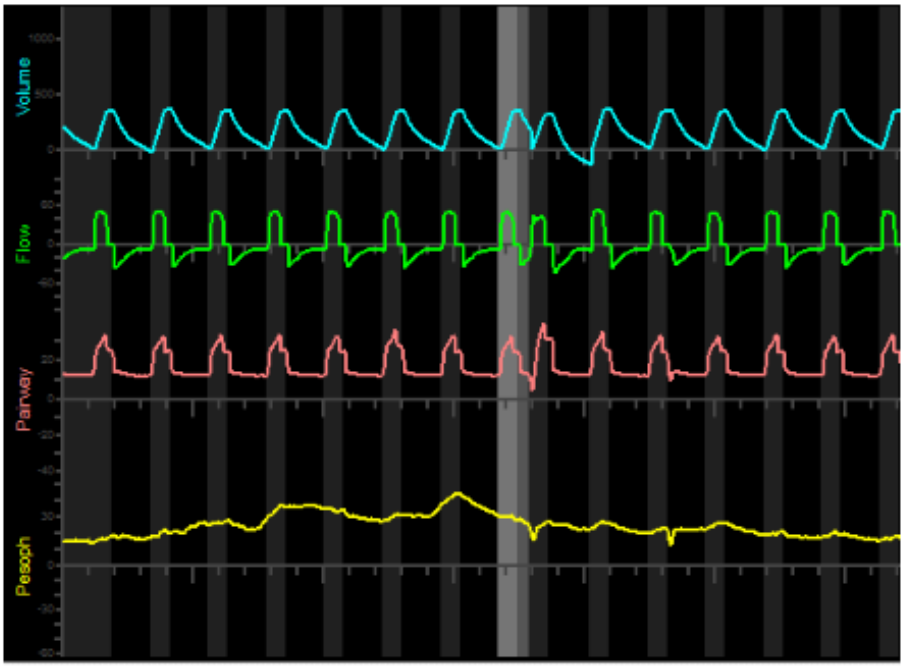


Figura 16. Imagen de zoom y tendencia de la respiración 367 del Centro médico 4.

La diferencia en las clasificaciones de los expertos se origina en si la respiración marcada y la siguiente son consideradas como un conjunto, o bien si se analizan por separado: se puede considerar que la primera es *reverse triggering* y la segunda no o viceversa. En el caso que se analice exclusivamente la respiración marcada y no se la considere como parte de un fenómeno que abarca la respiración posterior, se puede considerar que en esta respiración individual no hay asincronía y en la siguiente hay un *double triggering* o viceversa.

En general, puede verse que hay una diferencia de opiniones entre quienes consideran que *double triggering* es solo la primera respiración del par de doble disparo y quienes le asignan esta clasificación solo a la segunda. Por otro lado, si se asocia a la respiración marcada a la posterior, dado que la primera genera un *double triggering*, lo cual puede considerarse parte de un *reverse triggering*, se clasifican las dos respiraciones con una categoría que incluya al *double triggering*.

Por lo tanto, es importante señalar que en este caso el desacuerdo es relativo ya que se origina fundamentalmente en una diferencia de criterio para responder la encuesta y no en un desacuerdo respecto al fenómeno fisiológico (es decir, la asincronía). Es interesante contar con esta información, que antes de la utilización de la herramienta desarrollada no estaba disponible, para poder comenzar un debate en el ámbito de la salud respecto al criterio a utilizar para rotular respiraciones que se encuadren en este proceso fisiológico.

Una metodología para aumentar el consenso puede ser indicarles a los médicos al inicio de la encuesta de qué manera considerar las respiraciones que se encuadran en este caso, si analizarlas juntas o por separado. Ningún criterio sería erróneo, pero es importante que todos los evaluadores adopten el mismo ya que lo relevante sería aumentar la consistencia entre las consideraciones de los médicos para tener un Gold Standard de mayor calidad.

Respiración con bajo consenso N2

La segunda respiración con bajo consenso que se analizó se puede ver en Figura 17 y los resultados de las clasificaciones, en la Tabla 8. Se trata de la respiración 368 de uno de los archivos provenientes del Centro médico 4. Justamente, se trata de la respiración siguiente a la ya analizada (Respiración con bajo consenso N1) y los motivos de la gran discrepancia entre los evaluadores son los mismos a los mencionados en el inciso anterior.

Como se puede observar en la Tabla 8, 3 evaluadores (el 43% del total) responden “I don’t know”, es decir, indican que la respiración no es clara para ellos y no saben clasificarla. Otros 3 evaluadores (43% del total) seleccionan la categoría “Reverse triggering + Double triggering”. El evaluador restante (14% del total) indica que no hay asincronía presente.

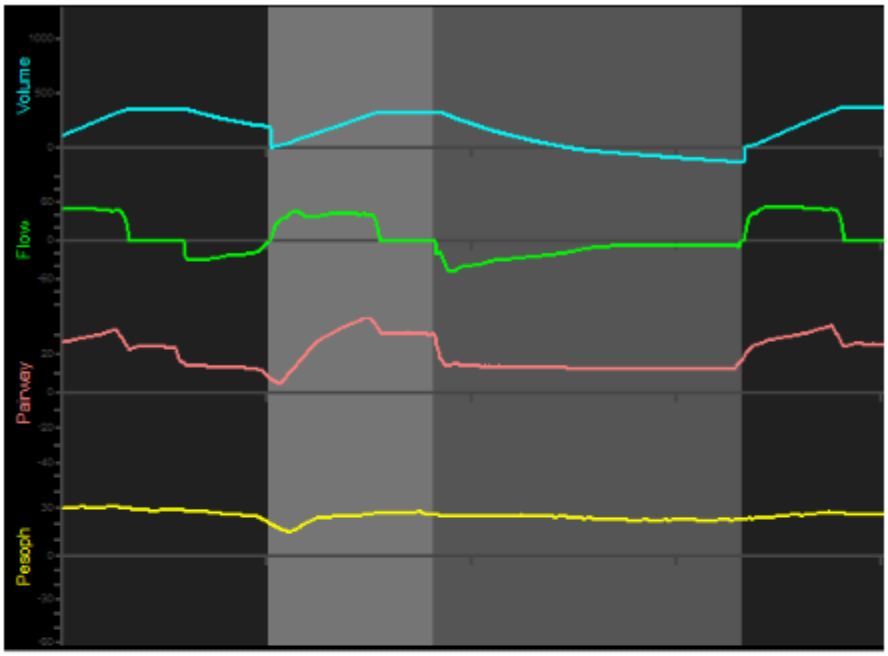
Evaluador	Respuesta
1	No asynchrony is present
2	Reverse triggering + Double triggering
3	I don’t know
4	Reverse triggering + Double triggering
5	I don’t know
6	I don’t know
7	Reverse triggering + Double triggering

Tabla 8. Resultados de las evaluaciones de expertos para la respiración 368 del Centro médico 4.

Es interesante poder comparar las respuestas para las respiraciones sucesivas de los distintos expertos. Debe destacarse que debido al orden aleatorio en el que se presentan las respiraciones en la plataforma desarrollada de encuestas, las mismas no se encontraban en forma sucesiva para los evaluadores al momento de responder.

El evaluador 1 clasifica la primera respiración como “Reverse triggering + Double triggering” y a la segunda como “No asynchrony is present”. El evaluador 2 clasifica a la primera respiración como “I don’t know” y a la segunda, “Reverse triggering + Double triggering”. El evaluador 4 clasifica tanto a la primera como a la segunda respiración como “Reverse triggering + Double triggering”. De forma similar, el evaluador 7 considera que la primera respiración es un “Double triggering (without reverse triggering)”, mientras que la segunda es un “Reverse triggering + Double triggering”. Por su parte, los evaluadores 3, 5 y 6 consideran que en la primera no hay asincronía (“No asynchrony is present”) mientras que no saben clasificar la segunda (“I don’t know”).

Zoom (the highlighted breath is the same as the one below)



Trend (the highlighted breath is the same as the one above)

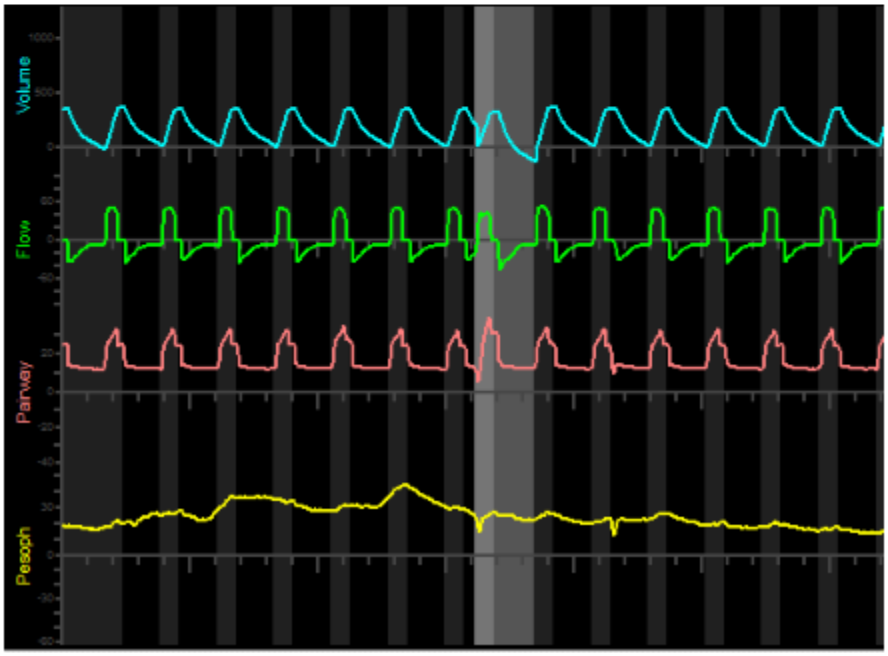


Figura 17. Imagen de zoom y tendencia de la respiración 368 del Centro médico 4.

El evaluador 1 rotula a las dos respiraciones asociadas de forma independiente. En particular, clasifica a la primera del par de respiraciones asociadas, como “Reverse triggering + Double triggering” mientras que en la segunda afirma que hay ausencia de asincronías. Podría considerarse al evaluador 2 dentro de este grupo teniendo en cuenta que asigna categorías distintas a las dos respiraciones del par: no sabe clasificar la primera, pero asigna la categoría “Reverse triggering + Double triggering” a la segunda. No obstante, por el hecho de que no detecta ninguna asincronía en la primera respiración, pero tampoco asegura que haya ausencia de asincronías, no se puede afirmar si clasifica a las respiraciones de modo independiente.

Por otro lado, los evaluadores 4 y 7 parecería que clasifican a ambas respiraciones del par en conjunto, destacando su pertenencia el mismo proceso fisiológico ya que ambos señalan la existencia de *double triggering* tanto en la primera como en la segunda. Se puede destacar que el evaluador 4 usa exactamente la misma clasificación para ambas respiraciones (“Reverse triggering + Double triggering”), mientras que el evaluador 7 utiliza “Double triggering (without reverse triggering)” para la primera y “Reverse triggering + Double triggering” para la segunda. Este criterio es válido ya que puede considerarse que el *reverse triggering* es generado por la primera respiración en la segunda. El hecho que considere que existe *double triggering* en ambas indicaría que las considera parte del mismo proceso.

Por último, no es claro si los evaluadores 3, 5 y 6 analizan ambas respiraciones del par como un conjunto o independientemente, dado que en la primera señalan que para ellos no existe asincronía, pero en la segunda no aclaran si detectan un *double triggering* (con o sin *reverse triggering*), aunque no descartan la posibilidad de que haya una asincronía en esa respiración.

Respiración con bajo consenso N3

La tercera respiración con bajo consenso analizada se puede ver en la

Figura 18 y los resultados de las clasificaciones se encuentran en la Tabla 9. Se trata de la respiración 664 de uno de los archivos provenientes del Centro médico 2.

Se puede observar que 3 de los evaluadores (43% del total) coinciden en que se trata de “Reverse triggering + Double triggering”, mientras que 2 evaluadores (28,5% del total) señalan que no hay asincronías (“No asynchrony is present”) y los 2 restantes (28,5%) no saben clasificar la respiración

("I don't know"). Dentro de los que sí saben clasificar la respiración, la opinión está dividida respecto de si se trata de un "Reverse triggering + Double triggering" o "No asynchrony is present".

Evaluador	Respuesta
1	Reverse triggering + Double triggering
2	Reverse triggering + Double triggering
3	I don't know
4	Reverse triggering + Double triggering
5	No asynchrony is present
6	No asynchrony is present
7	I don't know

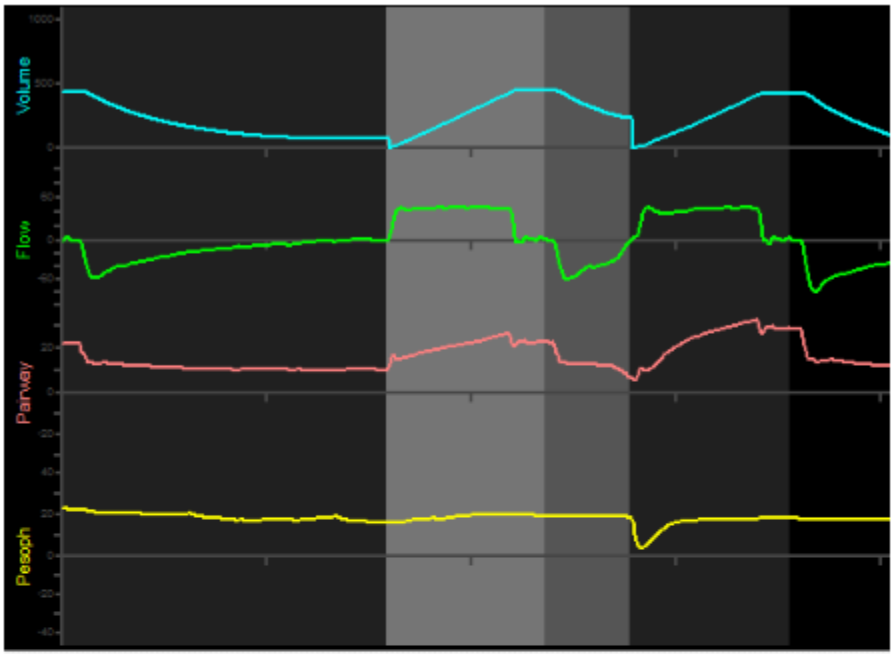
Tabla 9. Resultados de las evaluaciones de expertos para la respiración 664 del archivo Centro médico 2

De hecho, se trata de un caso muy similar a la respiración con bajo consenso N1 analizada, por ser la primera de un par de respiraciones asociadas que forman parte de un mismo proceso fisiológico. Las causas del desacuerdo son las mismas que en aquel caso. Parecería que se trata, esencialmente, de una diferencia de criterio para responder la encuesta y de rotulación, y no de un desacuerdo respecto al fenómeno fisiológico.

Se buscaron las respuestas para la respiración sucesiva a la analizada (la número 665 del Centro médico 2) para poder hacer una comparación en el criterio de evaluación de los distintos clasificadores. Las clasificaciones para la segunda respiración del par se encuentran en la Tabla 10.

Como se puede ver, para la respiración 665, el número máximo de clasificadores de acuerdo es de 4 (el 57% del total) que consideran que se trata de un "Reverse triggering + Double triggering". Los 3 clasificadores restantes (43%) consideran que no saben clasificar la respiración. Dentro de todos los que dicen saber clasificar la respiración, el 100% coinciden en que se trata de "Reverse triggering + Double triggering", pero no debe subestimarse el hecho que 3 expertos no hayan sabido clasificarla.

Zoom (the highlighted breath is the same as the one below)



Trend (the highlighted breath is the same as the one above)

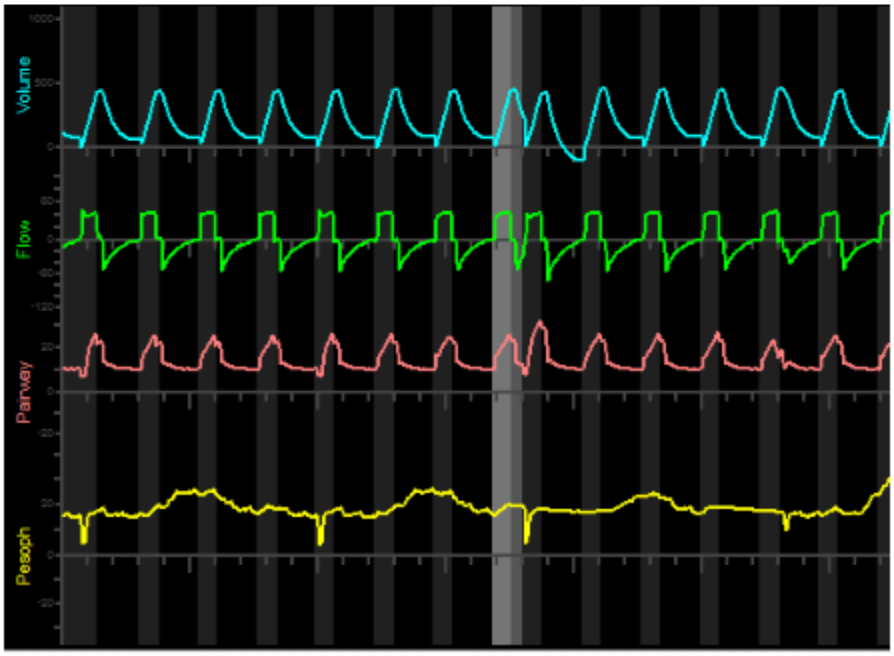


Figura 18. Imagen de zoom y tendencia de la respiración 664 del archivo Centro médico 2.

Evaluador	Respuesta
1	Reverse triggering + Double triggering
2	Reverse triggering + Double triggering
3	I don't know
4	Reverse triggering + Double triggering
5	I don't know
6	I don't know
7	Reverse triggering + Double triggering

Tabla 10. Resultados de las evaluaciones de expertos para la respiración 665 del Centro médico 2. Se trata de la respiración siguiente a la analizada (respiración con bajo consenso N3).

Los evaluadores 1, 2 y 4 analizan de forma conjunta a ambas respiraciones del par y las rotulan como "Reverse triggering + Double triggering". Por su parte, existe la posibilidad de que el evaluador 3 haya considerado a ambas respiraciones como parte del mismo fenómeno, pero no es claro cuál es su postura dado que clasifica a ambas como "I don't know", es decir, no sabe clasificar con certeza a ninguna de las dos respiraciones del par. Por último, los evaluadores 5 y 6 consideran que en la primera respiración del par hay ausencia de asincronías ("No asynchrony is present"), mientras que la segunda no la saben clasificar ("I don't know"). Por su parte, el evaluador 7 no sabe clasificar la primera respiración, pero clasifica a la segunda como "Reverse triggering + Double triggering".

Se hace una comparación entre los criterios de los evaluadores para las clasificaciones del par 367 y 368 del Centro médico 4 (analizado para las respiraciones con bajo consenso N1 y N2) y aquellas consideradas en esta sección correspondientes el par 664 y 665 del Centro médico 2. El objetivo es analizar cualitativamente la consistencia de los clasificadores consigo mismos.

El evaluador 1 no parecería ser consistente, ya que en el segundo caso clasifica a ambas respiraciones en conjunto mientras en el primer caso rotula a ambas respiraciones de forma separada (clasificaciones distintas). En cambio, el evaluador 4 es definitivamente consistente ya que en ambos casos rotula las dos respiraciones de la misma manera: "Reverse triggering + Double triggering". En cuanto los evaluadores 2 y 7, no se puede afirmar si son o no consistentes en la clasificación dado que, tanto el evaluador 2 para el par 367 y 368 del Centro médico 4 como el

evaluador 7 para el par 664 y 665 del Centro médico 2, no saben clasificar a la primera del par. Similar es el caso del evaluador 3 que para el segundo par no sabe clasificar ninguna respiración así que no se puede analizar la consistencia de su clasificación. Los evaluadores 5 y 6 en ambos casos rotulan la primera respiración como “No asynchrony is present” y la segunda como “I don’t know”, es decir, son consistentes consigo mismos en la clasificación de los dos pares de respiraciones, aunque no es claro si las analizan en conjunto o por separado.

Debe tenerse en cuenta que, como ya fue mencionado, el desacuerdo para este tipo de respiraciones es relativo ya que se origina en un desacuerdo (y en algunos casos, inconsistencia interna del evaluador consigo mismo) respecto al criterio para nomenclar al par de respiraciones y no respecto al proceso fisiológico. No obstante, conocer esta diferencia de opiniones e inconsistencias brinda la posibilidad de abrir un debate entre expertos para la futura unificación de criterio de clasificación de respiraciones. Esto ayudará a aumentar el acuerdo tanto general (inter-evaluadores) como interno (intra-evaluador).

Respiración con bajo consenso N4

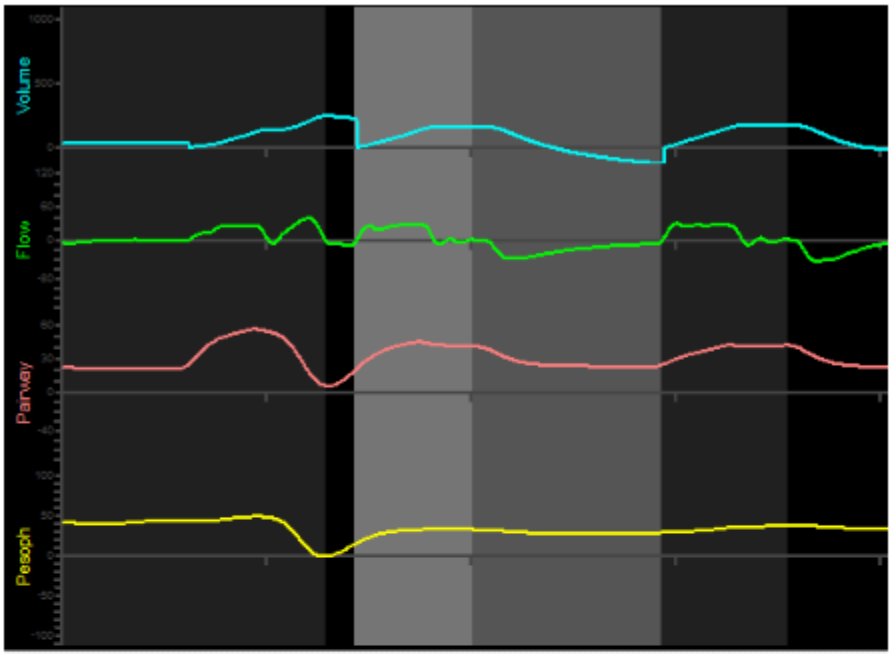
La cuarta respiración con bajo consenso analizada se puede ver en la

Figura 19 y los resultados de las clasificaciones, en la Tabla 11. Se trata de la respiración 6 de uno de los archivos provenientes del Centro Médico 1.

Evaluador	**Respuesta**
1	Reverse triggering + Double triggering
2	No asynchrony is present
3	I don’t know
4	Reverse triggering + Double triggering
5	I don’t know
6	I don’t know
7	Double triggering (without reverse triggering)

Tabla 11. Resultados de las evaluaciones de expertos para la respiración 6 del Centro Médico 1.

Zoom (the highlighted breath is the same as the one below)



Trend (the highlighted breath is the same as the one above)

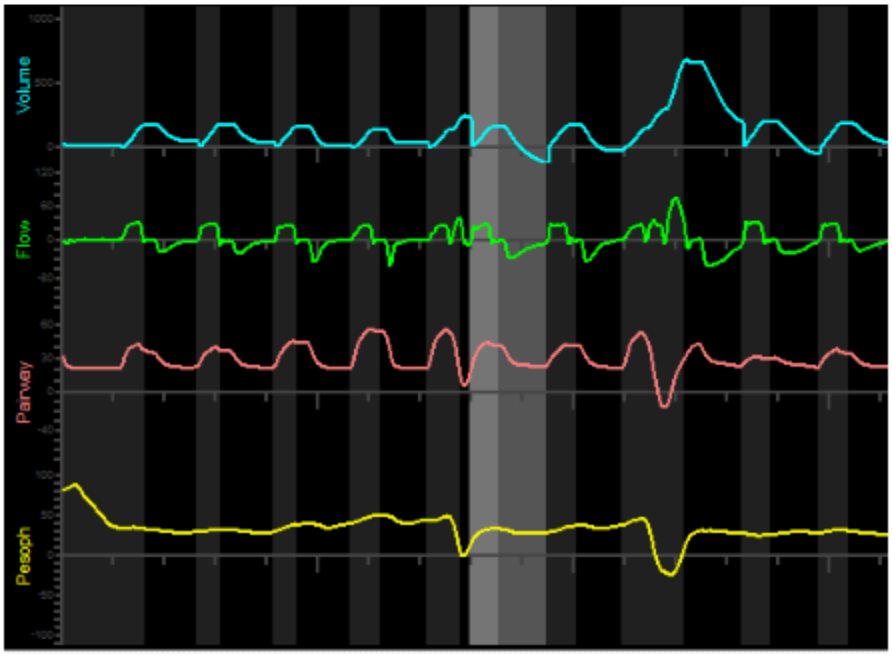


Figura 19. Imagen de zoom y tendencia de la respiración 6 del Centro Médico 1.

Se puede ver que la opinión se encuentra muy dividida. Tres evaluadores (43% del total) no saben clasificar la respiración ("I don't know"), otros dos clasificadores (29% del total) consideran que se trata de "Reverse triggering + Double trigginering", otro clasificador (14% del total) afirmó que se trata de "Double triggering (without reverse triggering)" y el restante (14% del total), que no hay asincronías ("No asynchrony is present"). Dentro de los clasificadores que sí saben clasificar la respiración, la mayoría consideran que se trata de un fenómeno relacionado a *double triggering* pero no hay acuerdo en si incluye *reverse triggering* o no. Por lo tanto, ni siquiera de este grupo de clasificadores se puede definir una categoría que sea la mayoritaria.

Este caso es un ejemplo de respiración compleja, en la que hay una falta de acuerdo entre los expertos respecto al fenómeno fisiológico que ocurre. No existe una posible metodología para aumentar el consenso entre evaluadores. Sin embargo, debe destacarse que estos casos de respiraciones confusas son siempre una minoría respecto al total de respiraciones.

Respiración con bajo consenso N5

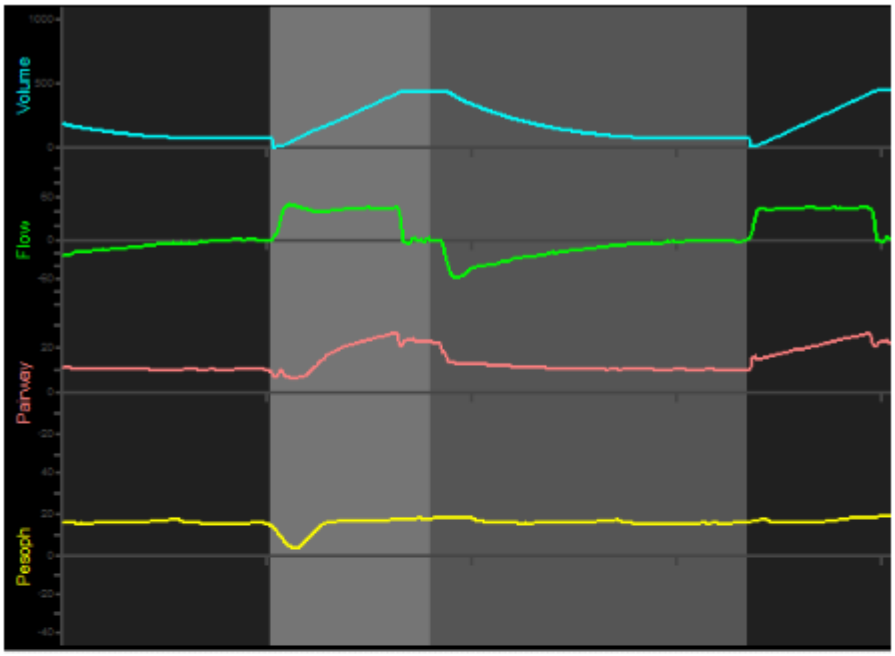
La quinta respiración con bajo consenso analizada se puede ver en la

Figura 20 y los resultados de las clasificaciones en la Tabla 12. Se trata de la respiración 661 de uno de los archivos provenientes del Centro médico 2.

Evaluador	Respuesta
1	No asynchrony is present
2	Long cycling
3	I don't know
4	Long cycling
5	I don't know
6	I don't know
7	Long cycling

Tabla 12. Resultados de las evaluaciones de expertos para la respiración 661 del Centro médico 2.

Zoom (the highlighted breath is the same as the one below)



Trend (the highlighted breath is the same as the one above)

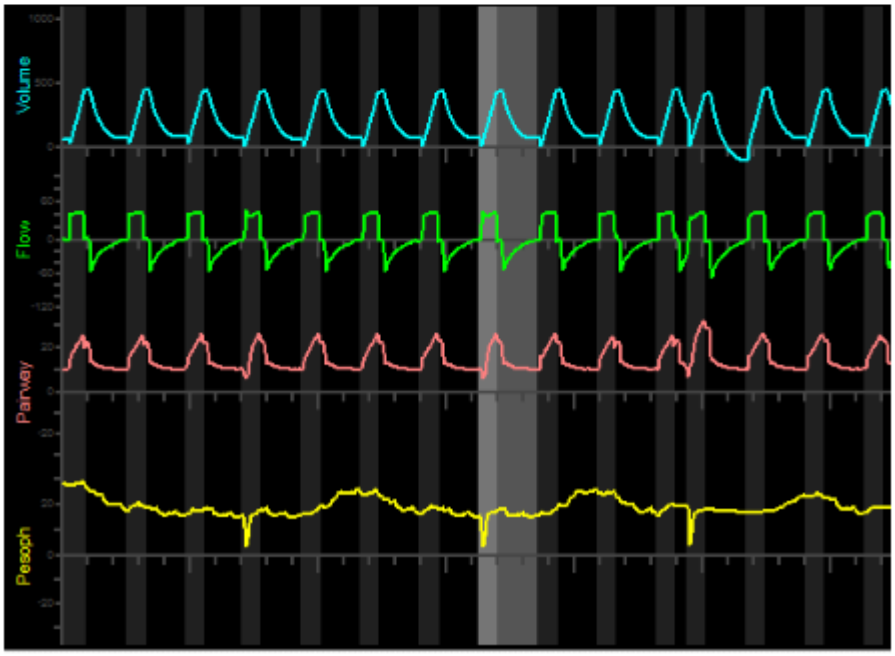


Figura 20. Imagen de zoom y tendencia de la respiración 661 del Centro médico 2.

Para esta respiración existe falta de consenso entre los evaluadores. Tres de ellos (43% del total) no supieron clasificar la respiración ("I don't know"), otros tres (43% del total) eligieron la categoría "Long cycling", mientras el evaluador restante (14% del total) no detectó ninguna asincronía presente ("No asynchrony is present"). Dentro de los que sí saben clasificar la respiración, la mayoría (75% respecto al grupo que clasifica) eligió la categoría "Long cycling". Sin embargo, se trata de una respiración confusa y con nivel bastante bajo de acuerdo dado que el 43% del total de evaluadores expertos no supo clasificar la respiración.

No existe un método que permita aumentar el consenso para esta respiración ya que el desacuerdo proviene de una interpretación distinta del proceso fisiológico, al igual que en el caso analizado para la respiración con bajo consenso N4. Es importante tener en cuenta que no es posible que haya un consenso absoluto entre los expertos para todas las respiraciones, sino que algunas tendrán bajo consenso y cada evaluador puede estar influenciado por su formación académica y contexto clínico. No obstante, debe señalarse, al igual que en el caso de la respiración con bajo consenso N4, que estas respiraciones con bajo nivel de acuerdo son una minoría respecto del total de respiraciones.

Conclusiones respecto a respiraciones con bajo consenso

En el proceso de armado de Gold Standard, es esperable encontrar una pequeña proporción de respiraciones que tengan un nivel relativamente bajo de acuerdo, es decir, en la que no haya un alto nivel de consenso entre los evaluadores. El origen del descuerdo incluye tanto factores de inconsistencia intra-evaluador como de desacuerdo entre evaluadores.

En el caso de ejemplo de clasificación de respiraciones en asincronías, debe señalarse que existen tanto respiraciones en las que los criterios de los evaluadores son distintos y existe una falta de acuerdo real, como otras en las que el desacuerdo es relativo ya que se origina fundamentalmente en una diferencia de criterio para responder la encuesta y no en un desacuerdo respecto al fenómeno fisiológico.

En particular, se pudo ver que es común que haya desacuerdo en la forma de nomenclar las respiraciones para los casos en las que se involucra un "Double triggering (without reverse triggering)" y "Reverse triggering + Double triggering", que son fenómenos que involucran dos respiraciones asociadas. Esta información, obtenida gracias a la utilización de la herramienta, es

valiosa para poder fomentar el debate entre especialistas respecto al criterio a utilizar para rotular dichas respiraciones, de modo que a futuro exista un criterio más generalizado que permita aumentar el consenso para clasificaciones de asincronías.

Existen alternativas para aumentar el acuerdo para una encuesta ya realizada, tal como la planteada por Krippendorf [29] que sugiere que se puede lograr un consenso en deliberaciones posterior a la codificación. Esto implicaría reunir a los expertos codificadores, posteriormente a que cada uno haya hecho su evaluación individual, mostrarles las respiraciones en las que hay discrepancias de manera que ellos puedan debatir e intercambiar opiniones. Sin embargo, Krippendorf aclara que los datos antes de dicha conciliación son datos de fiabilidad propiamente dichos y producen una fiabilidad mensurable. Además, señala que, aunque es razonable suponer que la conciliación de codificación posterior mejora la fiabilidad de los datos más allá de la fiabilidad de los datos generados por cualquier observador individual, este es un supuesto sin evidencia medible. La única fiabilidad publicable es la que se mide antes de la conciliación de desacuerdos y la fiabilidad de los datos después de este esfuerzo de reconciliación es discutible.

En este caso ejemplificador de clasificación de asincronías, no se realizó un trabajo post-encuesta de conciliación, y se decidió que la clasificación mayoritaria debía ser considerada la correcta. Sin embargo, se considera la posibilidad de considerar solo aquellas respiraciones (89 respiraciones) con un consenso mayor o igual a 5 evaluadores para el armado de un Gold Standard con mayor precisión.

Adicionalmente, se propone para una futura aplicación de la herramienta para crear Gold Standards en el área de asincronías, agregar como metodología para aumentar el consenso una indicación a los médicos al inicio de la encuesta que explique qué criterio se debe utilizar para clasificar las respiraciones que se encuadran el caso que incluye *reverse triggering* y *double triggering*.

Respiraciones con bajo consenso según paciente

Las señales respiratorias utilizadas para mostrar el funcionamiento de la herramienta generadora de encuestas son provenientes de pacientes internados en Unidad de Cuidados Intensivos con ventilación mecánica. Se trata de cinco pacientes de cinco centros de salud de la Ciudad Autónoma de Buenos Aires. Se evaluaron dos grupos de diez respiraciones sucesivas por paciente. Se elaboró

la Tabla 13 para averiguar en qué proporción se distribuyen las respiraciones con bajo nivel de consenso. Se consideran respiraciones con bajo nivel de acuerdo a aquellas en las que el número máximo de evaluadores de acuerdo es menor o igual a 4 (que representa un acuerdo en un porcentaje menor al 57,14% de todos los evaluadores). En total, hay once respiraciones con esas características que representan el 11% del total de respiraciones consideradas.

Centro médico del paciente	Respiraciones con bajo nivel de consenso
Centro médico 1	2 (18,2%)
Centro médico 2	5 (45,4%)
Centro médico 3	1 (9,1%)
Centro médico 4	2 (18,2%)
Centro médico 5	1 (9,1%)

Tabla 13. Tabla que muestra la distribución de las respiraciones con bajo nivel de consenso según el centro médico del que se obtuvo la señal.

Se puede observar que el 45,4% del total de respiraciones con bajo nivel de consenso para los expertos provienen de los registros del paciente del centro médico 2. Esto puede deberse al estado de gravedad del paciente o la configuración del ventilador.

También, es posible que el experto no marque la opción que consideró correcta en alguna respiración debido a la fatiga o el cansancio de realizar una tarea repetitiva. Esto aumenta el desacuerdo cuando, en realidad, es un error que proviene de realizar la tarea repetitiva de clasificar una cantidad de elementos relativamente grande (100 respiraciones en este caso de ejemplo) en forma consecutiva. Para este caso de ejemplo, se analizó la fatiga en la sección “Distribución del desacuerdo”, considerando la distribución de las respiraciones con bajo consenso respecto al orden de presentación de las mismas.

En una futura aplicación de la plataforma para el caso de asincronías respiratorias, se podría complementar dicho análisis estudiando las inconsistencias internas de los expertos, de acuerdo con la fatiga. Esto implicaría estudiar el cansancio del evaluador presentándole varias veces las mismas respiraciones con opiniones divergentes y analizando si existe una asociación entre la inconsistencia interna y el orden de presentación de las respiraciones.

Distribución del desacuerdo

La encuesta fue armada con 100 respiraciones, lo cual fue decidido a partir de una prueba preliminar con dos referentes. Sin embargo, se tiene en cuenta la posibilidad que, debido a la fatiga o el cansancio, el experto analice con menos atención las respiraciones a medida que avance la encuesta, aumentando con eso la probabilidad de que se equivoque.

Por lo tanto, es interesante analizar la distribución de las respiraciones con bajo nivel de consenso una vez que la encuesta fue realizada a los expertos, respecto a distintos grupos de respiraciones. Se consideraron 4 grupos de respiraciones, los cuales abarcan los siguientes números de respiraciones: 1-25, 26-50, 51-75 y 76-100. Los números de las respiraciones indican el orden en el que se le presenta la misma al usuario cuando está respondiendo la encuesta. También, se considera el acuerdo porcentual para estos cuatro grupos de respiraciones, que cuantifica el acuerdo para cada una de las respiraciones del grupo a partir de los pares posibles de respuestas. Los resultados se pueden observar en la Tabla 14.

Se analizó la distribución de los acuerdos porcentuales individuales correspondientes las respiraciones para analizar si dichos datos podían modelarse con una distribución normal. Se realizó la prueba de Shapiro–Wilk, diseñada para contrastar la normalidad de un conjunto de datos. Como era de esperar, observando la Tabla 14 en la cual se ven que los datos, que solo pueden ir de 0 a 100 (por definición del acuerdo porcentual), se concentran en los valores cercanos al extremo superior, la prueba dio un resultado negativo, por lo cual se verifica que los datos de acuerdo porcentual no tienen una distribución normal y no pueden modelarse de tal manera.

En primer lugar, se puede observar en la Tabla 14 ver que la mediana de los cuatro grupos de respiraciones es 100% lo cual muestra el alto nivel de acuerdo en todos ellos, a pesar de la existencia de respiraciones con bajo nivel de consenso específica.

Grupo de respiraciones	Respiraciones con bajo nivel de consenso	Acuerdo porcentual	Mediana	Primer cuartil	Tercer cuartil
1-25	2 (18,2%)	80,6 %	100	71,4	100
26-50	2 (18,2%)	83,4 %	100	71,4	100
51-75	2 (18,2%)	87,6 %	100	71,4	100
76-100	5 (45,4%)	75,8 %	100	50	100

Tabla 14. Distribución de las respiraciones con bajo nivel de consenso en los distintos grupos de respiraciones.

La cantidad de respiraciones con bajo nivel de consenso (donde el número máximo de expertos en acuerdo es igual o inferior a 4) es igual para los tres primeros grupos de respiraciones. Complementariamente, el porcentaje de acuerdo promedio sube ligeramente para el grupo de respiraciones 26-50 respecto al del grupo de respiraciones 1-25 (83,4% vs. 80,6%) y vuelve a subir en el grupo de respiraciones 51-75 (87,6%) respecto a los dos anteriores. De hecho, el mayor acuerdo porcentual se encuentra en este último grupo de respiraciones, indicando que hay un mayor consenso para estas respiraciones. Sin embargo, para el grupo de respiraciones 76-100 se encuentra la mayor concentración de respiraciones con bajo nivel de consenso (el 45,4% del total de respiraciones de bajo consenso están en este grupo, en comparación con el 18,2% para cada uno de los grupos anteriores) y el menor acuerdo porcentual (75,8%) que disminuyó 11,8% respecto al grupo de respiraciones inmediatamente anterior en orden (51-75). Esto podría significar que existió un cansancio de los expertos hacia el final de la encuesta, que pudo haber llevado a disminuir su concentración y, por lo tanto, el consenso.

Si se analiza la Tabla 14, la distribución de los datos de acuerdo porcentual por respiración se puede observar que si bien el grupo de (76-100) parecería ser distinto a los demás, dado que su primer cuartil es 50 (lo cual implica un mayor rango intercuartil) y su acuerdo porcentual promedio es el más bajo. Sin embargo, se hizo una prueba estadística no paramétrica y se pudo concluir que no existen diferencias estadísticamente significativas entre los diferentes grupos de respiraciones.

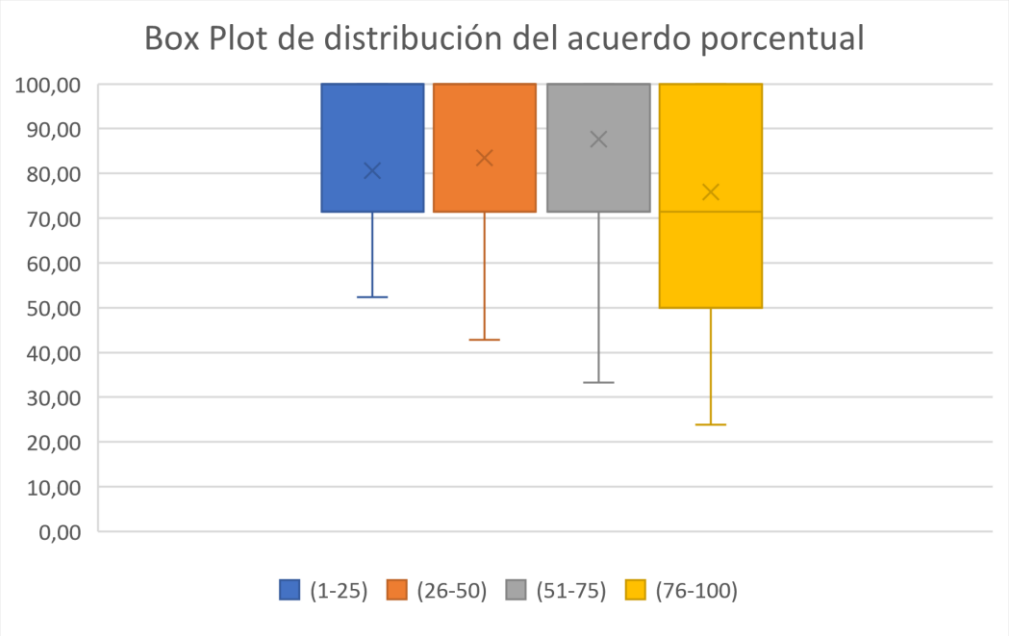


Figura 21. Box plot de la distribución del acuerdo porcentual por respiración respecto a distintos grupos de respiraciones

Dado que el consenso es esencial para la construcción de un *benchmark*, una posible opción para aumentar el consenso para un futuro Gold Standard en medicina respiratoria podría abarcar un total de 75 respiraciones teniendo en cuenta la concentración de respiraciones específicas con bajo consenso (aquellas con solo 3 o 4 evaluadores de acuerdo) y su menor acuerdo porcentual medio en comparación con los otros grupos. Sin embargo, dado que la diferencia en la distribución del acuerdo porcentual no es estadísticamente significativa para los distintos grupos de respiraciones, se consideraron las 100 respiraciones para la realización del Gold Standard en cuestión.

Modelado de acuerdo

Estadística Kappa

Hasta este momento se ha utilizado únicamente el denominado acuerdo porcentual para medir la fiabilidad inter-evaluadores para datos nominales, dado que es una herramienta muy utilizada en el área de estadística descriptiva por ser de fácil cálculo e interpretación. Sin embargo, tiene algunas desventajas ya que no considera el acuerdo que se realiza por casualidad. Por eso, una de las medidas de modelado (no ya de estadística descriptiva) más utilizadas es la estadística kappa.

La estadística kappa se trata de una medida de acuerdo corregida respecto al azar: compara el nivel observado de acuerdo (P_o), con el nivel de acuerdo esperado solo por azar (P_e). Los valores de kappa representan la proporción de acuerdo más allá del esperado por casualidad. Se distingue del acuerdo porcentual dado que, justamente, tiene en cuenta el acuerdo que se realiza por azar (es decir, considera la posibilidad que conjeturas entre evaluadores que sean congruentes entre sí).

Como la mayoría de las estadísticas de correlación, el kappa puede variar de −1 a +1 [33]. El valor de kappa es 1 cuando se produce un acuerdo perfecto entre los dos evaluadores, 0 cuando el acuerdo es igual al acuerdo esperado y negativo cuando el acuerdo es menor de lo esperado por azar.

La expresión general de una kappa [34] es:

$$k = \frac{P_o - P_e}{1 - P_e}$$

donde P_o es una probabilidad observada de acuerdo y P_e es la probabilidad de acuerdo esperada bajo ciertas restricciones de referencia que definen el acuerdo por casualidad.

Cohen formuló el kappa de Cohen [35] y el *weighted kappa* o kappa ponderado [36] que son válidos cuando hay únicamente dos evaluadores. Desde entonces se han realizado generalizaciones para casos de múltiples evaluadores. Entre ellos, se destacan las realizadas por Fleiss [37] y Fleiss et al [38], que tienen aplicación en el caso de diferentes conjuntos de múltiples evaluadores para cada elemento u objeto a clasificar (en el caso de ejemplo de clasificación de asincronías, es una respiración).

Para cada Gold Standard que se quiera construir, la estadística kappa que se calcula para medir el acuerdo debe elegirse considerando cuidadosamente el diseño de la encuesta en cuestión. La herramienta de creación de encuestas realizada para el caso de ejemplo de clasificación de asincronías generó una encuesta totalmente cruzada o *fully-crossed* en las que cada evaluador codifica a todas las respiraciones. Para este caso, además, la encuesta es respondida por siete evaluadores, que se encuadra en el caso de múltiples (se define como múltiples si son tres o más) evaluadores.

Para las encuestas totalmente cruzadas con evaluadores múltiples, se han desarrollado dos alternativas de kappa posibles. Por un lado, una propuesta formulada por Light et al [39] sugiere calcular kappa para todos los pares de codificadores y luego usar la media aritmética de estas estimaciones para proporcionar un índice de acuerdo general. Por su lado, Davies y Fleiss [40] plantean una derivación similar de la estadística kappa que usa el P(e) promedio entre todos los pares de codificadores para calcular una estadística similar a kappa, pero aplicable para codificadores múltiples. Dado que no existe un consenso generalizado acerca de qué kappa es una medida más adecuada, para medir el acuerdo en el caso de clasificación de asincronías se calcularon tanto el kappa de Davies y Fleiss como el Kappa de Light.

Kappa de Davies y Fleiss

La solución de Davies y Fleiss no se encuentra disponible en ningún paquete ni programa de estadística [31]. Por eso, en este trabajo, se implementó en Google Apps Script, dicha solución a partir de las ecuaciones desarrolladas por Davies y Fleiss en su trabajo [40].

Se integró dicha herramienta de forma directa con el *Spreadsheet* de respuestas generadas por la encuesta, es decir, al ser ejecutado se calcula el kappa de Davies y Fleiss tomando directamente la codificación de los expertos en la encuesta de clasificación. En el caso que se desee utilizar esta herramienta estadística para otros datos, se puede modificar en el script de forma sencilla la identificación (ID) del *Spreadsheet* con los datos a analizar.

En la formulación de kappa realizada por Davies y Fleiss, el error estándar asintótico se deriva bajo la hipótesis que las categorías de clasificación nominales son independientes. Por otro lado, las variables se ordenan en una disposición de doble entrada (sujeto vs. evaluador) y también se refiere

a este tipo de clasificaciones como *fully crossed* (cada uno de los evaluadores evalúa todos los sujetos), como es el caso de ejemplo de clasificación de respiraciones en asincronías.

En el trabajo publicado por Davies y Fleiss "Measuring Agreement for Multinomial Data" [40] se mencionan a modo de ejemplo posibles casos de aplicación de esta estadística kappa en particular, los cuales incluyen cuando un mismo conjunto de médicos asigna a cada uno de los pacientes en estudio una de las categorías de diagnóstico posibles, las cuales son mutuamente excluyentes. En la utilización de la herramienta de generación de Gold Standards que se realiza en este trabajo, cada médico asigna a cada respiración un tipo de asincronía, siendo las mismas mutuamente excluyentes. Lo que interesa es la precisión intrínseca de la clasificación.

Se debe suponer que cada elemento del conjunto (I) de unidades experimentales u elementos a clasificar (en este caso de ejemplo de asincronías respiratorias, se trata de cada una de las respiraciones) es clasificada en una de las C categorías mutuamente excluyentes (en esta aplicación, son clases de asincronías, la ausencia de las mismas o el desconocimiento de si hay o no asincronías), por cada uno de los miembros de un conjunto J de evaluadores (para este caso de ejemplo, los profesionales de la salud expertos en medicina respiratoria).

En la encuesta desarrollada para el caso de ejemplo de clasificación de asincronías, *I*=100 (100 respiraciones a ser clasificadas), *C*=10 (cantidad de categorías mutuamente excluyentes que se utilizan para rotular las respiraciones) y *J*=7 (cantidad de profesionales de la salud que realizan las clasificaciones).

El vector multinomial $X_{ij} = \left(X_{ij1}, X_{ij2}, \dots, X_{ijC}\right)'$ representa la clasificación resultante de la unidad o elemento (respiración) i (i = 1 , … , *I*) realizada por el evaluador j (j = 1, … , J). Por lo tanto, cada X_{ijc} (c = 1, … , C) toma el valor 1 o 0 y la $\sum_c X_{ijc} = 1$ para todo i y j, dado que el evaluador puede elegir una sola categoría entre todas para rotular la respiración. Por ejemplo, el evaluador 3 clasifica la respiración 74 como "I don't know", que en el vector de categorías generado (en el mismo orden que se presentan las opciones en la encuesta) es la categoría 10, por lo que el vector $X_{74\,3} = (0,0,0,0,0,0,0,0,0,1)'$.

Dado que para cada unidad u elemento hay un total de $\frac{1}{2}J(J-1)$ pares de clasificaciones, en este caso al ser J=7 evaluadores el total es de 21 pares. Para la unidad i, el número observado de pares que están de acuerdo es $\frac{1}{2}\sum_c Y_{ic}(Y_{ic}-1)$, donde $Y_{ic}=\sum_j X_{ijc}$ y representa el número de evaluadores clasificando la unidad i en la categoría c.

Para el caso de ejemplo de clasificación de asincronías, se representan las primeras 10 filas (para las primeras diez respiraciones i, objetos de clasificación) de la matriz Y en la Tabla 15. Cada elemento Y_{ic} representa cuántas veces fue rotulada la respiración i (en las filas) con la categoría c (en las columnas). Se encuentra la tabla completa que representa dicha matriz y las 100 respiraciones en el Anexo B.

i/c	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	0	0	0	0	0	0	0	7	0
2	0	0	0	1	0	0	0	2	3	1
3	0	0	0	0	0	0	0	7	0	0
4	0	0	0	0	0	0	0	7	0	0
5	0	0	0	1	0	0	0	6	0	0
6	0	0	0	1	0	0	0	2	1	3
7	0	0	0	0	0	0	2	0	5	0
8	0	0	0	0	0	0	0	0	6	1
9	0	0	0	0	0	0	0	0	7	0
10	1	0	0	0	0	0	6	0	0	0

Tabla 15. Diez (10) primeras filas de la Matriz Y calculada en por el algoritmo de kappa de Davies y Fleiss. Las filas representan las respiraciones y las columnas las categorías. Los elementos Yic indican cuántos evaluadores clasifican a la respiración i con la categoría c. La tabla completa se encuentra en el Anexo B.

Por ende, en la ecuación general de estadísticas kappa, teniendo en cuenta que $\sum_c Y_{ic}=\sum_j\sum_c X_{ijc}=J$ (en este caso de ejemplo de clasificación de asincronías, en el que hay 7 evaluaciones por respiración i, puede verificarse para cualquier i: $\sum_c Y_{ic}=\sum_j\sum_c X_{ijc}=7$) se debe reemplazar para el acuerdo observado:

$$P_o = \frac{1}{IJ(J-1)} \sum_i \sum_c Y_{ic} \left(Y_{ic} - 1\right) = \frac{1}{IJ(J-1)} \left(\sum_i \sum_c {Y_{ic}}^2 - IJ \right)$$

Para las respuestas obtenidas en la encuesta realizada, la probabilidad observada (P_o) es igual a 0,8195. Se puede ver que el acuerdo observado no es más que el acuerdo porcentual (81,9%).

Por otro lado, se define $\prod_j = (\prod_{j1}, \ldots, \prod_{jC})'$ como el vector de probabilidades multinomiales subyacente al vector X_{ij}. $\prod_{jc}$ se estima a partir de la proporción de todas las unidades (respiraciones) asignadas a la categoría c por el evaluador j:

$$P_{jc} = \frac{1}{I} \sum_i X_{ijc}$$

Reemplazando I=100 para este caso de ejemplo particular, se obtiene:

$$P_{jc} = \frac{1}{100} \sum_i X_{ijc}$$

Se representa en la Tabla 16 la representación de la matriz P, con sus elementos P_{jc} (por ejemplo, el evaluador 7 asigna 0,14 de todas sus clasificaciones a la categoría 7).

j/c	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,23	0,64	0,00
2	0,04	0,00	0,01	0,00	0,00	0,12	0,17	0,19	0,53	0,01
3	0,00	0,01	0,00	0,00	0,00	0,00	0,12	0,17	0,60	0,10
4	0,00	0,00	0,00	0,00	0,00	0,01	0,14	0,22	0,62	0,01
5	0,00	0,01	0,00	0,00	0,00	0,00	0,11	0,14	0,66	0,08
6	0,01	0,01	0,00	0,00	0,00	0,00	0,10	0,14	0,66	0,08
7	0,00	0,03	0,00	0,02	0,00	0,01	0,14	0,19	0,45	0,16

Tabla 16 Representación de la matriz P. Los elementos P_{jc} refieren a la proporción que cada evaluador j utiliza cada categoría c. Se calculan estas proporciones en base a todas las calificaciones que hace cada evaluador.

Si las clasificaciones de varios evaluadores son independientes para cada unidad, la probabilidad que un dado par (se abrevian como j y k) de evaluadores estén de acuerdo en la clasificación de un elemento aleatorio es $\sum_c \prod_{jc} \prod_{kc}$.

Por lo tanto, la probabilidad media de acuerdo de a pares debido al azar (es decir, que coincidan las evaluaciones cuando los evaluadores están meramente haciendo conjeturas) es:

$$\prod_e = \frac{2}{J(J-1)} \sum_{j>k} \sum_c \left(\prod_{jc} \prod_{kc} \right)$$

La probabilidad esperada (de que los evaluadores coincidan por azar) puede ser estimada de acuerdo a:

$$P_e = \frac{2}{J(J-1)} \sum_{j>k} \sum_c P_{jc} P_{kc}$$

$$P_e = \sum_c \bar{P}_c^{\,2} - \frac{2}{J(J-1)} \sum_c \sum_j \left(P_{jc} - \bar{P}_c\right)^2$$

En esta fórmula de P_e se tiene en cuenta que $\bar{P}_c = \sum_j {P_{jc}}/{J}$, que es la proporción de clasificaciones en la categoría c. En particular, para este ejemplo de clasificación de asincronías, $\bar{P}_c = [0{,}007\ ,\ 0{,}009\ ,\ 0{,}001\ ,\ 0{,}004\ ,\ 0{,}000\ ,\ 0{,}009,\ 0{,}130\ ,\ 0{,}183\ ,\ 0{,}594\ ,\ 0{,}063\ ]$.

El P_e resultante de dicha estimación para este ejemplo de asincronías es de 0,40.

Finalmente, reemplazando P_e y P_o en la ecuación de kappa general, se obtiene que el kappa de Davies y Fleiss para los resultados obtenidos es igual a 0,70, como se ve en la siguiente ecuación:

$$k = \frac{P_o - P_e}{1 - P_e} = \frac{0{,}8195 - 0{,}4060}{1 - 0{,}4060} = 0{,}70$$

Por otro lado, si se consideran para los mismos cálculos únicamente las 89 respiraciones en las que hay un acuerdo de 5, 6 o 7 evaluadores (aquellas en las que hay un acuerdo en un porcentaje superior al 71,4% de todos los evaluadores para esta encuesta), es decir, excluyendo las 11 respiraciones con bajo nivel de consenso, con el fin de obtener un Gold Standard de mayor precisión,

se obtiene, reemplazando $P_e = 0{,}46$ y $P_o = 0{,}88$ un kappa de 0,78. Como es esperable, el kappa de acuerdo aún teniendo en cuenta la posibilidad de que haya consenso por azar (representado por el componente de P_e en la fórmula de kappa), es superior cuando se eliminan las 11 respiraciones con bajo nivel de consenso.

Kappa de Light

Light et al [39] formula una propuesta diferente a la de Davies y Fleiss para calcular el kappa entre múltiples evaluadores para estudios totalmente cruzados. Se trata de una derivación para múltiples evaluadores del kappa de Cohen [35], una de las medidas de consenso para el caso de dos evaluadores. Se propone calcular el kappa de Cohen para todos los pares de codificadores y luego usar la media aritmética de estas estimaciones para proporcionar un índice de acuerdo general.

Para evaluar el caso de asincronías respiratorias, se utilizó la implementación de kappa de Light del paquete "irr" (inter-rater reliability) desarrollado en lenguaje R. Las respuestas obtenidas de los evaluadores pueden descargarse del *Spreadsheet* de modo xls y ser transformadas a una matriz en R de forma directa.

Si se considera la totalidad de los datos sin distinguir entre respiraciones con bajo y alto nivel de consenso (es decir, las 100 respiraciones analizadas) para calcular el kappa de Light, se obtiene un kappa de 0,70.

Por otro lado, si se tienen en cuenta únicamente las 89 respiraciones en las que hay un acuerdo de 5, 6 o 7 evaluadores, excluyendo las 11 respiraciones con bajo nivel de consenso se obtiene un kappa de 0,79.

Al igual que en el caso de kappa de Davies y Fleiss, el kappa de Light es superior cuando se eliminan las 11 respiraciones con bajo nivel de acuerdo. Por lo tanto, considerar únicamente las 89 respiraciones con alto nivel de consenso (aquellas en las que hay un acuerdo en un porcentaje superior al 71,4% de todos los evaluadores para esta encuesta) implicaría tener un Gold Standard de mayor precisión.

Interpretación de Kappa

Existe un gran debate respecto cuál es el nivel de kappa necesario para justificar la fiabilidad de los datos y esto es de gran interés para construir un Gold Standard. En particular, se desea conocer el rango de kappa que se corresponde con un determinado nivel de acuerdo [31]. El objetivo final de la prueba de fiabilidad es asegurar que la no fiabilidad sea insignificante para justificar un análisis de los datos y, para esta aplicación en partcular, para justificar la generación de un Gold Standard con dichos datos (en este caso, se puede ejemplificar con los kappas obtenidos para la encuesta de asincronías respiratorias).

Uno de los criterios más adoptados es el planteado por Landis y Koch [41]. En su trabajo, se establecen pautas para interpretar los valores de la estadística kappa. Si el valor de kappa se encuentra entre 0.0 y 0.2, el acuerdo es leve; entre 0.21 y 0.40 el acuerdo es justo; entre 0.41 y 0.60 el acuerdo es moderado, entre 0.61 y 0.80, el acuerdo es sustancial; y por último, valores de kappa entre 0.81 y 1.0 indican acuerdo casi perfecto o perfecto.

Sin embargo, el uso de estos puntos de corte cualitativos ha sido debatido y no existe conenso absoluto sobre estos límites. Krippendorff planteó en su trabajo de 1980 [29] una interpretación más conservadora que sugiere que las conclusiones no deben ser aceptadas para las variables con valores de kappa menores a 0,67, mientras que pueden establecerse tentativamente para los valores de kappa entre 0,67 y 0,80, y solo pueden ser definitivas las conclusiones para valores superiores a 0.80.

En la práctica, los coeficientes de kappa por debajo de los valores de corte conservadores de Krippendorff a menudo se mantienen en los estudios de investigación, y Krippendorff ofrece estos valores de corte basados en su propio trabajo en el análisis de contenido, al tiempo que reconoce que las estimaciones de IRR aceptables variarán dependiendo de los métodos de estudio y la pregunta de investigación.

Para el caso de la encuesta de asincronías paciente-respirador, tanto el kappa de Davies y Fleiss como el kappa de Light calculados teniendo en cuenta el totalidad de las respiraciones es de 0,70 . Estos resultados indican que el acuerdo es sustancial según los lineamientos de rangos establecidos por Landis y Koch, por lo cual es válido el desarrollo de un Gold Standard a partir de estos datos.

Incluso, puede afirmarse que las conclusiones obtenidas pueden establecerse tentativamente de acuerdo al criterio más conservador de Krippendorf. Esto último implicaría que tentativamente se podría armar un Gold Standard que abarque las 100 respiraciones.

Considerando las 89 respiraciones con alto nivel de consenso (aquellas en las que el acuerdo es de 5, 6 o 7 evaluadores), se obtiene un kappa Davies y Fleiss de 0,78 y un kappa de Light de 0,79. Dichos kappas también implican que el acuerdo es sustancial en función de los rangos planteados por Landis y Koch y que las conclusiones obtenidas pueden establecerse tentativamente de acuerdo al criterio de Krippendorf. Sin embargo, aunque se encuentran en el mismo rango, se debe destacar que estos valores de kappa indican un acuerdo mayor a aquel que muestran los kappas obtenidos para el total de respiraciones. De hecho, el kappa de Davies y Fleiss se encuentra a 0,02 y el Kappa de Light se encuentra a 0,01 del rango considerado por Landis y Koch como acuerdo perfecto y para el cual se pueden establecer conclusiones definitivas según los límites más conservadores de Krippendorf.

Es interesante observar que el kappa de Davies y Fleiss y el kappa de Light, si bien son diferentes en su cálculo y no existe un acuerdo respecto de cuál es la medida más apropiada, dan resultados muy similares. Cuando se calculan en base a la totalidad de las respiraciones son iguales (aproximando respecto al segundo decimal) mientras que cuando se computan en base a 89 respiraciones difieren solo en 0,01 (el kappa de Davies y Fleiss es de 0,78 y el kappa de Light es de 0,79). Por lo tanto, la utilización de ambos kappas como medidas de acuerdo permiten obtener las mismas conclusiones, que son consistentes con las obtenidas con el acuerdo porcentual. El hecho de que el acuerdo porcentual indique un alto nivel de acuerdo y que ambos kappas calculados, considerando las respiraciones totales y las respiraciones con alto nivel de consenso (que representan el 89% de las totales), indican un acuerdo sustancial (de acuerdo con Landis y Koch) e incluso permitan establecer conclusiones tentativamente (según el criterio más conservador de Krippendorf) muestra cuan robusto es el Gold Standard que se puede generar a partir de la herramienta desarrollada.

Análisis de experiencia de usuario

Se hizo un doble análisis de la experiencia de usuario. En primer lugar, se analizó la experiencia de los usuarios de la plataforma, los cuales quieren generar un Gold Standard. También, se consideró la experiencia de los evaluadores al responder la encuesta que fue creada por la herramienta, en el caso de ejemplo de clasificación de asincronías.

La experiencia de los usuarios de la herramienta de generación de encuestas es buena dado que la misma es sencilla de usar y rápida. El usuario debe disponer de una base de señales que desee clasificar y definir las categorías independientes que quiera utilizar para generar el Gold Standard. El tiempo total del armado del *Spreadsheet* que se utiliza como datos de entrada del *script* y la ejecución de este es de aproximadamente 30 minutos, que es considerado rápido. Además, el análisis de las respuestas una vez obtenidas las evaluaciones de los expertos también es fácil de realizar. Se pueden obtener las clasificaciones automáticamente en el *Spreadsheet* de respuestas, del cual se pueden descargar o bien ejecutar sobre el *Spreadsheet* las funciones que se consideren adecuadas para medir el consenso.

En cuanto a la experiencia del evaluador completando la encuesta que genera la herramienta, se consultaron opiniones a los expertos involucrados en la encuesta de ejemplo de clasificación de asincronías. Se realizó una encuesta en la que se le solicitó a los expertos que califiquen su experiencia cualitativamente como “muy buena”, “buena”, “regular” o “mala”. El 85,7% de los evaluadores (6 evaluadores) la calificaron como “muy buena”, mientras que el 14,3% (1 evaluador), como “buena”. Estos resultados son muy alentadores e indican una usabilidad muy satisfactoria.

Hubo quienes respondieron la encuesta en formato móvil y quienes respondieron en formato PC y hubo comentarios sobre la apropiada visualización de la encuesta en ambos casos, lo cual muestra cuan efectiva que es la herramienta y la elección de un entorno web. Si bien un experto señaló algunos aspectos que pueden optimizarse en futuras versiones de la plataforma, tal como la ampliación de la escala de las imágenes (más allá del zoom que se puede realizar manualmente con la computadora o el celular), no opinó que esto sea un obstáculo para la utilización de la misma y destacó como útil la solución adoptada de observar dos escalas de tiempo para cada elemento.

En cuanto a la posible fatiga del experto durante el proceso de llenado de la encuesta, previamente a la realización de la encuesta se hizo una prueba preliminar con dos referentes, en base a la cual se decidió utilizar 100 respiraciones. El tiempo estimado de duración para completar la encuesta generada en este caso de ejemplo es de 30 a 40 minutos. Posteriormente a la realización de la encuesta final a los expertos, se analizó la distribución del acuerdo porcentual en diferentes grupos de respiraciones, para ver si la distribución del acuerdo guardaba coherencia con la prueba preliminar.

Los resultados de la sección “Distribución del desacuerdo” indican que el mayor acuerdo porcentual se encuentra en el grupo de respiraciones 51-75 (87,6%) mientras que se reduce en el grupo de que abarca las respiraciones 76 a 100 (75,8%). En este último grupo también se concentra la mayor proporción de respiraciones con bajo nivel de consenso (45,4%). Sin embargo, se hizo una prueba estadística y se pudo concluir que no existe una diferencia estadísticamente significativa en la distribución de los acuerdos porcentuales individuales, entre los diferentes grupos de respiraciones analizados (1-25, 25-50, 51-75 y 76-100).

Complementariamente, en la encuesta a los expertos sobre su experiencia de usuario, un evaluador destacó que dedicó una mayor cantidad de tiempo para responder en las primeras respiraciones y aceleró el ritmo de respuesta hacia el final. También, otro experto comentó que si bien no dedicó sustancialmente menos tiempo a medida que avanzaba en la encuesta, por el cansancio dejó de mirar las escalas numéricas para las distintas curvas y se centró, principalmente, en la forma de onda para hacer la clasificación.

Dado que en ambas devoluciones se mencionó la fatiga y que los datos cuantitativos indican una mayor tasa de desacuerdo en las últimas 25 respiraciones, se considera válida la opción de reducir la cantidad de respiraciones de 100 a 75 para futuras utilizaciones de la herramienta desarrollada en el área de asincronías respiratorias, ya que aumentaría el consenso (aunque no de forma significativa estadísticamente). Sin embargo, teniendo en cuenta que la diferencia en la distribución de los acuerdos porcentuales no es estadísticamente significativa para los distintos grupos de respiraciones, se consideró una decisión igualmente apropiada utilizar las 100 respiraciones para este Gold Standard.

Debe destacarse que no existe un número adecuado de elementos a clasificar que sirva para todas las encuestas, sino que el usuario debe definirlo para cada Gold Standard de manera individual. Entre los factores a considerar, se debe estimar el tiempo requerido para clasificar cada elemento y el nivel de atención necesario. Además, se recomienda realizar para cada caso particular una prueba preliminar con varios expertos para comprobar si la cantidad de elementos a clasificar seleccionada a priori es apropiada para la encuesta final.

La herramienta desarrollada tuvo una gran facilidad y rapidez de uso para el usuario interesado en generar un Gold Standard y, además, generó una encuesta en la cual el experto evaluador tuvo también una buena experiencia de usuario. En conjunto, la satisfactoria experiencia de usuario tanto del usuario de la plataforma como del evaluador que realizó las clasificaciones muestran la utilidad y el gran potencial de adopción de la herramienta desarrollada para facilitar el proceso de construcción de Gold Standards.

Conclusiones

Debido a la amplitud de uso de la herramienta desarrollada para facilitar el proceso de generación de Gold Standards, se decidió mostrar su funcionamiento y utilidad mediante un ejemplo de uso. Se utilizó la herramienta para generar una base señales respiratorias nomencladas en asincronías paciente-ventilador. La población elegida consistió en pacientes de UCI con ventilación mecánica y SDRA, por ser un área en la cual sería relevante contar con dicho Gold Standard para mejorar los resultados clínicos.

Su utilización fue sencilla dado que la encuesta, en base a la cual se obtuvieron los datos para generar el Gold Standard, se generó automáticamente al ejecutar el *script* y sus respuestas se almacenaron en un *Spreadsheet,* formato que permitió un directo procesamiento posterior. Los únicos datos de entrada que debió proveer el usuario son la base de señales fisiológicas a nomenclar y la clasificación deseada (en este caso, asincronías paciente-ventilador). Se programó la modificación del software FluxView para que permita generar imágenes de señales respiratorias, que para su utilización como usuario no requiere ningún conocimiento previo de informática. Por eso, este mismo software podría ser empleado como herramienta por cualquier usuario para crear otro Gold Standard de este tipo de señales.

La plataforma desarrollada, en base a Google Apps Script, permitió la generación de una encuesta totalmente cruzada con 100 respiraciones a clasificar, que fue el número elegido para este caso de ejemplo. En futuras aplicaciones, cada vez que se utilice la herramienta para generar un Gold Standard se debe definir la cantidad de elementos a clasificar para cada señal fisiológica y clasificación particular. Entre los factores que deben considerarse para tomar esta decisión se pueden mencionar el nivel de análisis y tiempo que requiere clasificar cada elemento. Para el caso de ejemplo de clasificación de asincronías, el análisis de la distribución de las respiraciones con bajo nivel de consenso de acuerdo con el número de respiración indicaría que 75 respiraciones podrían haber tenido un mayor consenso que 100 presumiblemente debido la fatiga del experto. Sin embargo, se mantuvieron las 100 respiraciones en este caso dado que un análisis estadístico del acuerdo porcentual por respiración indicó que las diferencias del acuerdo porcentual entre los grupos de respiraciones no son estadísticamente significativas.

En cuanto a los datos obtenidos de las clasificaciones de los evaluadores, los resultados de los análisis de concordancia para este caso de ejemplo son muy alentadores ya que demostró un elevado nivel de acuerdo entre los expertos. En base a un análisis descriptivo, existe un muy alto nivel de consenso en 89% de las respiraciones analizadas: coinciden en su clasificación 5, 6 o 7 evaluadores que representan el 71,4%, 85,7% o 100% de todos los expertos.

Considerando la totalidad de respiraciones (100), el acuerdo porcentual es de 81,9% y tanto el kappa de Fleiss como el kappa de Light fueron de 0,70. Si se tienen en cuenta únicamente las 89 respiraciones en las que hay un acuerdo superior al 71,4% de los evaluadores, el acuerdo porcentual es de 88,0% y el kappa de Davies y Fleiss y el kappa de Light son 0,79 y 0,78 respectivamente. Todos los valores anteriores indican un acuerdo sustancial y para el cual se pueden establecer conclusiones tentativamente. Como era de esperar, si se consideran únicamente las 89 respiraciones con alto consenso para armar el Gold Standard, los valores del acuerdo porcentual y de ambas estadísticas kappa se acercan a ser considerados acuerdo perfecto y para los cuales se pueden establecer conclusiones definitivamente. Estos resultados son coherentes e indican que el acuerdo es elevado en todos los casos, ya sea considerando las 100 respiraciones totales o las 89 en las que hay alto consenso. Por lo tanto, los datos recolectados en la encuesta a los expertos podrían ser utilizados construir un Gold Standard, lo cual evidencia la utilidad de la herramienta desarrollada para el caso de asincronías.

En cuanto a la experiencia de los usuarios de la herramienta de generación de encuestas, la experiencia fue buena dado que la herramienta fue encontrada como sencilla y rápida de usar (se estima que el tiempo total de generación de una encuesta es de 30 minutos). Complementariamente, respecto a la experiencia del profesional de la salud completando la encuesta generada por la herramienta, la mayoría de los evaluadores (6 evaluadores que representan el 85,7% del total) la calificaron como “muy buena” y solo 1 evaluador (que representa el 14,3% restante) como “buena”. Asimismo, destacaron la buena visualización en PC y formato móvil. Si bien se señalaron algunos aspectos que pueden optimizarse, ningún experto opinó que haya tenido dificultades para hacer la clasificación. En conjunto, la buena experiencia de usuario tanto del usuario que quiere armar la base de señales nomencladas como del experto que realiza la

clasificación en la encuesta, muestra la utilidad y el potencial de adopción de la herramienta en grupos de profesionales de la salud para facilitar el proceso de construcción de Gold Standards.

En un contexto en el que es necesario desarrollar una gran cantidad de Gold Standards: para las diferentes señales fisiológicas, patologías y grupos poblacionales, es relevante contar con una herramienta que simplifique la generación de dichas bases de señales nomencladas, como la que se ha desarrollado en este trabajo. Los buenos resultados de consenso, junto con la facilidad y rapidez de uso de la herramienta y la buena experiencia de usuario tanto para aquel que desea generar el Gold Standard como para el experto que responde la encuesta, demuestran el gran potencial de la herramienta desarrollada para simplificar el proceso de generación de Gold Standards, no sólo para el área respiratoria sino para cualquier señal fisiológica. Estos Gold Standards pueden ser utilizados para validar algoritmos de clasificación y detección automática y, de este modo, optimizar el funcionamiento de los sistemas de soporte de decisiones y del avance de la medicina digital. La herramienta desarrollada puede facilitarles el camino a los investigadores en estas nuevas tecnologías.

Se alcanzó satisfactoriamente el objetivo de mínima establecido al inicio del proyecto, que consistió en el desarrollo de una herramienta informática multidispositivo (apta tanto para PC como para móvil) que permite a los profesionales de la salud evaluar señales para clasificarlas y cuyos datos permitan facilitar la construcción de un Gold Standard. Asimismo, mediante el ejemplo de uso de clasificación de asincronías, también se pudo alcanzar el objetivo de máxima. Gracias a la colaboración de expertos en ventilación mecánica en este ejemplo de uso, quienes clasificaron las señales, se pudo demostrar la utilidad de la herramienta para obtener un Gold Standard específico. En este caso de uso, se trató un Gold Standard para señales respiratorias clasificadas en asincronías paciente-respirador.

La experiencia al realizar este Trabajo Final de Carrera fue muy formativa, dado que involucró la coordinación de diferentes aspectos que caracterizan a la Bioingeniería, tanto diseño y programación de herramientas informáticas (desde el software que genera las imágenes hasta el *script* que crea el formulario y las herramientas que calculan estadísticas de kappa, algunas no implementadas anteriormente como el kappa de Davies y Fleiss), como análisis estadístico y de experiencia de usuario y, por primera vez en la Carrera, interacción directa con profesionales de la

salud y otros ingenieros de diferentes especialidades. Siendo Bioingeniería una carrera dedicada a los desafíos del futuro en materia de tecnología aplicada a la medicina, fue muy enriquecedor poder tener contacto y desarrollar una herramienta que tiene como usuarios a profesionales de la salud y que puede facilitar la investigación en medicina digital.

Bibliografía

[1] Food & Drug Administration U.S., “Clinical and Patient Decision Support Software - Draft Guidance for Industry and Food and Drug Administration Staff,” 2017.

[2] L. Blanch *et al.*, “Asynchronies during mechanical ventilation are associated with mortality,” *Intensive Care Med.*, vol. 41, no. 4, pp. 633–641, 2015.

[3] A. W. Thille, P. Rodriguez, B. Cabello, F. Lellouche, and L. Brochard, “Patient-ventilator asynchrony during assisted mechanical ventilation,” *Intensive Care Med.*, vol. 32, no. 10, pp. 1515–1522, Sep. 2006.

[4] L. Vignaux *et al.*, “Patient–ventilator asynchrony during non-invasive ventilation for acute respiratory failure: a multicenter study,” *Intensive Care Med.*, vol. 35, no. 5, pp. 840–846, May 2009.

[5] T. Pham, I. Telias, T. Piraino, T. Yoshida, and L. J. Brochard, “Asynchrony Consequences and Management,” *Crit. Care Clin.*, vol. 34, no. 3, pp. 325–341, Jul. 2018.

[6] M. A. Gentile, K. D. Hargett, D. Chipman, J. Villar, and R. Kacmarek, “Cycling of the mechanical ventilator breath.,” *Respir. Care*, vol. 56, no. 1, pp. 52–60, Jan. 2011.

[7] S. Parthasarathy, A. T. Jubran, and M. J., “Cycling of Inspiratory and Expiratory Muscle Groups with the Ventilator in Airflow Limitation,” *Am. J. Respir. Crit. Care Med.*, vol. 158, no. 5, pp. 1471–1478, Nov. 1998.

[8] R. D. Branson, “Patient-Ventilator Interaction: The Last 40 Years,” *Respir. Care*, vol. 56, no. 1, pp. 15–24, Jan. 2011.

[9] A. W. Thille, P. Rodriguez, B. Cabello, F. Lellouche, and L. Brochard, “Patient-ventilator asynchrony during assisted mechanical ventilation,” *Intensive Care Med.*, vol. 32, no. 10, pp. 1515–1522, Sep. 2006.

[10] C. S. Sassoon and G. T. Foster, “Patient-ventilator asynchrony.,” *Curr. Opin. Crit. Care*, vol. 7, no. 1, pp. 28–33, Feb. 2001.

[11] E. Garofalo *et al.*, “Recognizing, quantifying and managing patient-ventilator asynchrony in invasive and noninvasive ventilation,” *Expert Rev. Respir. Med.*, vol. 12, no. 7, pp. 557–567, Jul. 2018.

[12] D. C. Chao, D. J. Scheinhorn, and M. Stearn-Hassenpflug, “Patient-ventilator trigger asynchrony in prolonged mechanical ventilation.,” *Chest*, vol. 112, no. 6, pp. 1592–9, Dec. 1997.

[13] R. D. Branson, T. C. Blakeman, and B. R. Robinson, “Asynchrony and Dyspnea,” *Respir. Care*, vol. 58, no. 6, pp. 973–989, Jun. 2013.

[14] M. de Wit, K. B. Miller, D. A. Green, H. E. Ostman, C. Gennings, and S. K. Epstein, “Ineffective triggering predicts increased duration of mechanical ventilation *,” *Crit. Care Med.*, vol. 37, no. 10, pp. 2740–2745, Oct. 2009.

[15] D. C. Chao, D. J. Scheinhorn, and M. Stearn-Hassenpflug, “Patient-ventilator trigger asynchrony in prolonged mechanical ventilation.,” *Chest*, vol. 112, no. 6, pp. 1592–9, Dec. 1997.

[16] L. Vignaux *et al.*, “Performance of noninvasive ventilation algorithms on ICU ventilators during pressure support: a clinical study,” *Intensive Care Med.*, vol. 36, no. 12, pp. 2053–2059, Dec. 2010.

[17] B. R. Robinson *et al.*, “Patient-Ventilator Asyncrhony in a Traumatically Injured Population,” *Respir. Care*, 2013.

[18] D. Colombo *et al.*, “Efficacy of ventilator waveforms observation in detecting patient–ventilator asynchrony*,” *Crit. Care Med.*, vol. 39, no. 11, pp. 2452–2457, Nov. 2011.

[19] I. I. Ramirez *et al.*, “Ability of ICU Health-Care Professionals to Identify Patient-Ventilator Asynchrony Using Waveform Analysis,” *Respir. Care*, vol. 62, no. 2, pp. 144–149, Feb. 2017.

[20] F. Longhini *et al.*, “Efficacy of ventilator waveform observation for detection of patient-ventilator asynchrony during NIV: a multicentre study.,” *ERJ open Res.*, vol. 3, no. 4, Oct. 2017.

[21] M. Dres, N. Rittayamai, and L. Brochard, “Monitoring patient–ventilator asynchrony,” *Curr. Opin. Crit. Care*, vol. 22, no. 3, pp. 246–253, Jun. 2016.

[22] H. Tokioka *et al.*, “The effect of breath termination criterion on breathing patterns and the work of breathing during pressure support ventilation.,” *Anesth. Analg.*, vol. 92, no. 1, pp. 161–5, Jan. 2001.

[23] D. Tassaux, J.-B. Michotte, M. Gainnier, P. Gratadour, S. Fonseca, and P. Jolliet, “Expiratory trigger setting in pressure support ventilation: from mathematical model to bedside.,” *Crit. Care Med.*, vol. 32, no. 9, pp. 1844–50, Sep. 2004.

[24] K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial.,” *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.

[25] D. J. Putka, H. Le, R. A. McCloy, and T. Diaz, “Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability.,” *J. Appl. Psychol.*, vol. 93, no. 5, pp. 959–981, 2008.

[26] K. Vaporidi *et al.*, “Clusters of ineffective efforts during mechanical ventilation: impact on outcome,” *Intensive Care Med.*, vol. 43, no. 2, pp. 184–191, Feb. 2017.

[27] D. Ciliska, N. Cullum, and A. Dicenso, “The fundamentals of quantitative measurement,” *Evid. Based. Nurs.*, vol. 2, no. 4, pp. 100–101, Oct. 1999.

[28] M. Pagano and K. Gauvreau, *Fundamentos de Bioestadística*. México D.F.: Thompson Learning, 2001.

[29] Klaus Krippendroff, *Content analysis : an introduction to its methodology*. Beverly Hills: Sage publications, 1980.

[30] M. L. McHugh, “Interrater reliability: the kappa statistic.,” *Biochem. medica*, vol. 22, no. 3, pp. 276–82, 2012.

[31] K. A. Hallgren, “Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial.,” *Tutor. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.

[32] John Uebersax, “Statistical Methods for Diagnostic Agreement,” 2015. [Online]. Available:

http://john-uebersax.com/stat/agree.htm. [Accessed: 19-May-2019].

[33] S. R.C. Bajpai, H.K. Chaturvedi, “Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods,” *J. Indian Acad. Appl. Psychol.*, vol. 41, no. 3, pp. 20–27, 2015.

[34] K. L. Posner, P. D. Sampson, R. A. Caplan, R. J. Ward, and F. W. Cheney, “Measuring interrater reliability among multiple raters: An example of methods for nominal data,” *Stat. Med.*, vol. 9, no. 9, pp. 1103–1115, Sep. 1990.

[35] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[36] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.,” *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.

[37] J. L. Fleiss, “Measuring nominal scale agreement among many raters.,” *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[38] J. L. Fleiss, J. C. Nee, and J. R. Landis, “Large sample variance of kappa in the case of different sets of raters.,” *Psychol. Bull.*, vol. 86, no. 5, pp. 974–977, 1979.

[39] R. J. Light, “Measures of response agreement for qualitative data: Some generalizations and alternatives.,” *Psychol. Bull.*, vol. 76, no. 5, pp. 365–377, 1971.

[40] M. Davies and J. L. Fleiss, “Measuring Agreement for Multinomial Data,” *Biometrics*, vol. 38, no. 4, p. 1047, Dec. 1982.

[41] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data.,” *Biometrics*, vol. 33, no. 1, pp. 159–74, Mar. 1977.

Anexo A

Ventilación mecánica

En la VM existen diversos modos ventilatorios, con el patrón de flujo, volumen y presión en la vía aérea del aire entregado al paciente. No existe un modo ventilatorio ideal, sino que se debe elegir uno en base a la patología del paciente y los objetivos del tratamiento.

La clasificación clásica de los modos ventilatorios los divide en dos grandes grupos, de acuerdo con la variable sobre la cual se basa el mecanismo: volumen o presión. Se puede conocer la información del paciente en aquella señal que no es directamente controlada por el respirador.

A continuación se mencionan algunos de los modos más utilizados en VM, algunos de los cuales ya fueron mencionados en la Tabla 1.

Modos controlados por volumen

En los modos volumétricos, un volumen específico de gas es entregado para llegar al Vt configurado. En consecuencia, Vt es consistente, se debe seleccionar el flujo pico o tiempo de inspiración y se debe configurar un patrón de flujo de gas. La variable resultante es la presión, la cual puede variar de respiración a respiración, y depende para cada paciente de la *compliance* y resistencia de la vía aérea. En cuanto a los distintos modos dentro de los volumétricos, se pueden mencionar: Volume *Controlled Ventilation* (VCV), *Volume Controlled Ventilation/Assist-Controlled* (VCV/AC), *Sinchronized Intermittent Obligatory Ventilation* (SIMV) e *Intermittent Mandatory Ventilation* (IMV).

En cuanto a VCV, el volumen corriente y la tasa de flujo para cada respiración son definidas por el médico en el respirador, mientras que la presión de la vía aérea depende de la *compliance* y resistencia del paciente. La tasa respiratoria también es definida y las respiraciones se ciclan por tiempo, por lo que Ti, Te y la tasa I:E son fijos.

Para VCV/AC, el volumen corriente y la tasa de flujo para cada respiración también son configuradas por el médico en el respirador. También, se define una mínima tasa respiratoria (*back up mandatory rate*). A diferencia del modo CMV en el cual todas las respiraciones son cicladas por tiempo (el "controlled" de su nombre significa que son controladas), en Assist CMV permite que el ciclado lo

realice el paciente (el “assist” de su nombre se refiere a que son respiraciones del paciente asistidas por el ventilador y no controladas por este). El paciente puede demandar una cantidad de respiraciones superior a la mínima establecida por la mínima tasa respiratoria, siempre que se supere el umbral de sensibilidad para asistencia preconfigurado. En general, las respiraciones suelen ser una combinación de respiraciones asistidas y controladas.

En el caso de SIMV hay respiraciones mecánicas (cicladas por volumen con un volumen y flujo preconfigurados por el médico) y espontáneas. El número de respiraciones mecánicas también es pre-definido por el médico y las mismas pueden ser cicladas por el paciente (asistidas) o cicladas por tiempo (controladas). El número de respiraciones espontáneas es determinado por el paciente y varía de minuto a minuto.

VMI es similar a SIMV excepto que las respiraciones mecánicas son brindadas en intervalos regulares sin tener en cuenta la actividad del paciente (solo ciclado por tiempo).

Modos controlados por presión

En los modos que tienen como target la presión, se configura un pico máximo de presión de la vía aérea, lo cual limita a su vez la presión alveolar. Por lo tanto, los targets son la presión de la vía aérea y alveolar, Vt puede variar de respiración a respiración y la entrega de flujo de aire es variable. Existen diferentes modos principales, denominados *Pressure Control Ventilation* (PCV), *Pressure Control Ventilation/Assist Control* (PCV/AC) y *Pressure Suport Ventilation* (PSV).

PCV es un modo de soporte ventilatorio en el cual el médico controla el tiempo de inspiración y la presión de vía aérea durante la inspiración. Los flujos entregados son los necesarios para lograr la presión inspiratoria deseada, y el volumen es dependiente de la interacción de estas configuraciones y la *compliance* y resistencia del sistema respiratorio del paciente. Las respiraciones controladas por presión pueden ser cicladas por tiempo (“controlled”) o cicladas por el paciente (“assisted”). Si se utiliza este modo (PCV) para prolongar el tiempo de inspiración de modo que la tasa I:E sea inversa se denomina PCIRV (Pressure Controlled Inversed Relation Ventilation).

También existe un modo de PCV/AC, cuyo modo de operación es similar al VCV/AC solo que el target en este caso es la presión. La operación básica es que cada respiración es iniciada por el paciente,

pero con una mínima tasa respiratoria que establece que, si el paciente no inicia una respiración durante un cierto tiempo, el ventilador automáticamente inicia una respiración auxiliar (controlada). Se debe configurar el nivel de presión, el tiempo inspiratorio, la mínima tasa respiratoria y la sensibilidad.

Por su parte, PSV provee un nivel de presión inspiratoria configurada por el médico, para cada esfuerzo del paciente. Por lo tanto, puede ser utilizado en cualquier modo de respiración espontánea. PS se inicia cuando la caída de presión en la inspiración cumple el umbral de sensibilidad para asistencia. El flujo entonces ingresa en el circuito del paciente para aumentar la presión de soporte establecida. El flujo continúa en proporción a la demanda del paciente hasta que el caudal respiratorio disminuya a un 25% respecto al pico inicial de flujo. En este punto, el soporte de presión termina. El paciente interactúa con la presión otorgada para determinar el tiempo de inspiración, el flujo y el volumen corriente. Se puede afirmar que las curvas de presión, volumen y flujo durante PSV son iguales a las de PCV/AC, con la distinción que el mecanismo que termina la respiración en el caso de PCV/AV es cuando el tiempo de inspiración configurado se agota, mientras que en PSV depende del flujo.

Anexo B

El algoritmo de Davies y Fleiss para calcular la estadística kappa involucra en sus cálculos a la matriz Y, cuyos elementos Yic representan la cantidad de veces que cada respiración i (en las filas) fue clasificado con la categoría c (en las columnas).

A continuación, la representación de la matriz Y obtenida en el algoritmo desarrollado y aplicado para el caso de 100 respiraciones (i) y 10 categorías (c).

i/c	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0	0	0	0	0	0	0	0	7	0
2	0	0	0	1	0	0	0	2	3	1
3	0	0	0	0	0	0	0	7	0	0
4	0	0	0	0	0	0	0	7	0	0
5	0	0	0	1	0	0	0	6	0	0
6	0	0	0	1	0	0	0	2	1	3
7	0	0	0	0	0	0	2	0	5	0
8	0	0	0	0	0	0	0	0	6	1
9	0	0	0	0	0	0	0	0	7	0
10	1	0	0	0	0	0	6	0	0	0
11	0	0	0	0	0	0	0	0	7	0
12	0	0	0	0	0	0	0	0	7	0
13	0	0	0	0	0	0	0	7	0	0
14	0	0	0	0	0	0	0	0	7	0

15	0	0	0	0	0	0	0	0	7	0
16	0	1	0	0	0	0	0	0	6	0
17	0	0	0	0	0	0	0	7	0	0
18	0	0	0	0	0	0	0	0	7	0
19	1	0	0	0	0	0	0	0	6	0
20	0	0	0	0	0	0	0	7	0	0
21	0	1	0	0	0	0	0	0	6	0
22	0	0	0	0	0	0	0	0	6	1
23	0	0	0	0	0	0	0	0	6	1
24	0	2	0	0	0	0	0	5	0	0
25	0	0	0	0	0	0	0	7	0	0
26	0	0	0	0	0	0	6	1	0	0
27	0	0	0	0	0	0	0	0	6	1
28	0	0	0	0	0	0	0	0	6	1
29	0	0	0	0	0	0	0	0	7	0
30	0	1	0	0	0	1	0	0	5	0
31	0	0	0	0	0	0	0	7	0	0
32	0	0	0	0	0	0	7	0	0	0
33	0	0	0	0	0	0	0	5	0	2
34	0	0	0	0	0	0	0	7	0	0
35	0	0	0	0	0	0	0	0	7	0

36	0	0	0	0	0	0	0	0	7	0
37	0	0	0	0	0	0	0	0	6	1
38	0	0	0	0	0	0	0	0	7	0
39	0	0	0	0	0	0	1	0	6	0
40	0	0	0	0	0	0	0	0	6	1
41	0	0	0	0	0	0	4	0	0	3
42	0	0	0	0	0	0	0	7	0	0
43	0	0	0	0	0	0	0	0	7	0
44	0	0	0	0	0	0	0	0	6	1
45	0	0	0	0	0	0	0	0	7	0
46	0	0	0	0	0	0	0	0	7	0
47	0	0	0	0	0	0	0	0	7	0
48	0	0	0	0	0	0	0	0	7	0
49	0	0	0	0	0	0	7	0	0	0
50	0	0	0	0	0	0	4	0	3	0
51	0	0	0	0	0	0	0	0	7	0
52	0	1	0	0	0	0	0	0	6	0
53	0	0	0	0	0	0	7	0	0	0
54	0	0	0	0	0	0	0	0	7	0
55	0	0	0	0	0	0	7	0	0	0
56	1	0	0	0	0	0	0	0	6	0

57	0	0	0	0	0	0	0	0	7	0
58	0	0	0	0	0	0	0	0	7	0
59	0	0	0	0	0	0	0	0	7	0
60	0	0	0	0	0	0	0	0	6	1
61	0	0	0	0	0	0	0	7	0	0
62	0	0	0	0	0	0	7	0	0	0
63	0	0	0	0	0	0	0	0	7	0
64	0	0	0	0	0	0	0	0	6	1
65	0	0	0	0	0	1	0	0	6	0
66	0	0	0	0	0	0	0	0	7	0
67	0	0	0	0	0	0	0	0	7	0
68	0	0	0	0	0	0	0	0	7	0
69	0	0	0	0	0	0	0	0	7	0
70	0	0	0	0	0	1	0	1	4	1
71	0	0	0	0	0	0	7	0	0	0
72	0	0	0	0	0	0	0	0	7	0
73	0	0	0	0	0	0	0	0	6	1
74	0	0	1	0	0	0	0	0	4	2
75	0	0	0	0	0	0	7	0	0	0
76	0	0	0	0	0	0	0	0	7	0
77	0	0	0	0	0	0	0	5	0	2

78	0	0	0	0	0	3	0	0	1	3
79	0	0	0	0	0	0	0	0	6	1
80	0	0	0	0	0	0	0	7	0	0
81	0	0	0	0	0	0	0	7	0	0
82	0	0	0	0	0	0	0	0	7	0
83	1	0	0	0	0	0	0	0	6	0
84	0	0	0	0	0	0	1	0	5	1
85	0	0	0	0	0	0	0	0	7	0
86	0	0	0	0	0	0	1	0	6	0
87	0	0	0	0	0	0	0	3	1	3
88	0	0	0	0	0	0	0	0	7	0
89	0	0	0	0	0	0	0	3	2	2
90	0	0	0	0	0	0	0	0	6	1
91	0	0	0	0	0	0	0	0	7	0
92	0	0	0	0	0	0	0	0	7	0
93	0	0	0	0	0	0	0	0	6	1
94	0	0	0	0	0	0	7	0	0	0
95	0	0	0	0	0	0	3	0	0	4
96	1	0	0	0	0	0	0	0	6	0
97	0	0	0	0	0	0	0	0	7	0
98	0	0	0	0	0	0	0	7	0	0

99	0	0	0	0	0	0	7	0	0	0
100	0	0	0	0	0	0	0	4	0	3